Maynooth University
National University of Ireland Maynooth

MAYNOOTH UNIVERSITY

DOCTORAL THESIS

JULY 23, 2024

GeoPrice: The development of an efficient, rapidly-updating, mix-adjusted median property price index model using stratified geospatial matching

---

Author:
Robert MILLER

Supervisor:
Dr. Phil MAGUIRE

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy

in the

Department of Computer Science
Faculty of Science & Engineering

Declaration of Authorship

I, Robert MILLER, declare that this thesis titled, "*GeoPrice*: The development of an efficient, rapidly-updating, mix-adjusted median property price index model using stratified geospatial matching" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: Robert Miller

Date: 26/08/2024

MAYNOOTH UNIVERSITY

Abstract

Faculty of Science & Engineering

Department of Computer Science

Doctor of Philosophy

***GeoPrice*: The development of an efficient, rapidly-updating, mix-adjusted median property price index model using stratified geospatial matching**

by Robert MILLER

The topic of house price index modelling is one which is central to a significant number of market stakeholders; governments, central banks, homeowners and businesses, among others. The impact which property price indices have on inflation, economic growth and policy-making are profound, yet the methodology and processes behind the generation of these statistics tools are rather opaque.

National statistical agencies will typically use one of two de-facto standard methods for modelling the housing market, those being hedonic regression and repeat-sales. While these methods bring with them distinct advantages, they also suffer from significant drawbacks. One of the most problematic of the these is the volume of rich property data required by the model. These data requirements often necessitate the use of non-public data sources, usually acquired through privileges as a government agency. As such, it is difficult for end-users of these statistics to verify the veracity, reliability and accuracy of the results.

Furthermore, these intensive data requirements induce a typical lag to publication in excess of two months. As a result, not only homeowners and businesses, but even policy-makers are operating on stale information, which is a substantial limitation given the critical influence exerted by the housing market on so many facets of the economy.

Our proposal is a novel, geospatially stratified house price index model which can be computed automatically on publicly available datasets. The algorithm does not require additional, privately-held attribute data for each property, nor does it necessitate a great deal of statistical expertise to implement, maintain and interpret, as the existing standards do. In this thesis, we will outline our methodology and demonstrate the performance of the index, initially on the Irish property market.

Following an initial study on Irish sale transactions, the model is extended to a database of asking prices for homes online, thus demonstrating the flexibility of the approach. This illustrates the accessibility of the model to operate on a variety of data sources. Finally, our algorithm will be employed to create a property price index for the United Kingdom, where the public dataset of sale transactions is significantly more plentiful. The results of this demonstrate that our index is not only as good as the official hedonic regression model produced by the ONS in the UK, but far exceeds the smoothness and noise reduction achieved by said model, while

maintaining a month-to-month correlation in excess of 85%. Moreover, our proposal achieves this with a lag time from data publication in the order of hours, rather than weeks, as per the ONS house price index.

Acknowledgements

This project and thesis would not have been possible without the extraordinary help of my supervisor, Phil. Phil has always been there to help advise me, open doors for me and sit around late for a chat when I needed it. I will always be grateful for his kind support and the time he has given me throughout my period studying under him, both through my undergraduate and postgraduate degrees.

I would also like to thank my mother, Adele. Life wasn't always easy growing up for my mother, myself and my two brothers. She sacrificed everything for the three of us and worked so hard to give us anything and everything she could, at her own expense. I can never thank her enough for all that she has done for me throughout my life. I wouldn't be where I am today without the blood, sweat and tears that she put in, through thick and thin. My two brothers, Paul and Karl, and my extended family, particularly my grandmother Bebe, also have been a major support, always encouraging me and taking interest in my career.

Finally, I must thank my partner Patrick, who has put up with endless nights of me staying up late and has never complained once. I appreciate you always being there to support me, celebrate with me and cheer me up when I'm feeling down.

Contents

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

CPI	**C**onsumer **P**rice **I**ndex
CSO	**C**entral **S**tatistics **O**ffice
EPC	**E**nergy **P**erformance **C**ertificate
HPI	**H**ouse **P**rice **I**ndex
MSM	**M**ean **S**pike **M**agnitude
NN	**N**earest **N**eighbour
ONS	**O**ffice for **N**ational **S**tatistics
PPR	**P**roperty **P**rice **Register**
PSRA	**P**roperty **S**ervices **R**egulatory **A**uthority
RPPI	**R**esidential **P**roperty **P**rice **I**ndex
SCB	**S**mallest **C**ommon **B**ucket
SE-trees	**S**et **E**numeration trees
VOA	**V**aluation **O**ffice **Agency**

Chapter 1

Introduction

1.1 The role of price indices

Price indices are a critical tool in economics and policy-making; their primary purpose being to measure changes in the cost of goods and services over time. The most common formulations of price indices are the Laspeyres, Paasche and Fisher indices, each of which will be outlined in this section, along with their distinct strengths and weaknesses.

While these indices can be applied to any basket of goods and services (such as the raw materials used by producers, for example), perhaps the most important use of price indices is in deriving the rate of consumer price inflation (CPI): the annual percentage change in a comprehensive, representative weighted basket of consumer goods and services.

Measurement of the *CPI* is of interest to many stakeholders, for a wide variety of reasons. Central bank policy makers keenly monitor the consumer inflation rate in order to assess the state of the economy and inform their monetary policy. The rate of inflation running too high erodes the purchasing power of consumers, leading to a potential slump in demand and economic recession. Deflation, which is when the annual inflation rate runs below zero, is also highly problematic, as it encourages consumers to wait for non-essential goods and services to become cheaper before purchasing them.

Price indices, such as the CPI, are often linked to contracts for pay and provision of services. For example, public sector employment contracts and social benefits,

such as pensions, are frequently directly linked to the rate of consumer price inflation, i.e. are *inflation-linked*. As such, each year, the payments made under these contracts are automatically up-rated by the inflation rate derived from the index. Similarly, many long-term service contracts, such as rental agreements, are inflation-linked, resulting in annual rent increases in-line with inflation.

These applications of economic price indices are also important in allowing accurate measurement of other economic indicators. The change in *nominal gross domestic product* (GDP), which is the monetary market value of all final goods and services produced in a country (typically within a one-year period), is not an accurate measure of economic growth without adjusting for inflation (Brezina, 2011). Suppose, for example, that inflation is running at a 2% annual rate, while the gross value of all produced goods and services is also 2% higher than a year prior. Given that, in this case, the increase in gross value is **entirely** due to price inflation, the change in *real gross domestic product* is zero, indicating no economic growth.

The use of price indices to compute the rate of inflation, along with applying inflation-adjustments to economic statistics, informs policy makers in how they should adjust monetary policy in order to keep inflation in their target range and stimulate or dampen economic growth. However, the price index methodology used to calculate it varies internationally, with different nations adopting different index forms in their methodology.

1.1.1 The Laspeyres index

The *Laspeyres* index was one of the earliest methodologies used in measuring the changes in price of goods and services and is still widely used today. For example, the consumer price indices produced by both the *Office for National Statistics* in the United Kingdom and the *Bureau of Labor Statistics* in the United States employ a Laspeyres-style index in their respective methodologies.

This index formulation measures the change in price of a basket of goods and services by comparing the sampled prices in a given time period relative to the prices of the same basket observed in a pre-determined *base period*. For example, if the base period is chosen as January 2020, the price index would be constructed by taking the quotient of the weighted-average price of the goods sampled in each subsequent

month with the weighted-average of the base period prices, where the weights remain fixed to those set in the base period. Formally:

$$L_t = \frac{\sum\limits_{c \in C} p_{c,t} * q_{c,0}}{\sum\limits_{c \in C} p_{c,0} * q_{c,0}}$$

where:

L_t is the Laspeyres price index level in period t

C is the collection of goods and services in the price basket

$p_{c,t}$ is the price of component c in period t

$p_{c,0}$ is the price of component c in the base period

$q_{c,0}$ is the quantity, or weight, of component c in the base period

The Laspeyres index is a popular choice in measuring consumer price trends due to its simplicity; the weights of the components need only be observed in the base period and can be updated periodically, as desired. As a result, the data requirements of this type of index are lower and thus, it can be computed and published in a more efficient and timely manner than indices which require more frequent re-weighting.

One key drawback of this decision to fix the consumption weights to the base period is known as *substitution bias*. In cases where the price of a particular component increases substantially, a portion of consumers are likely to switch to a cheaper, competing product. The static nature of the weights used in the Laspeyres index means that this effect is not captured within the price index and, as such, the index does not respond accurately to real price-demand market dynamics. This typically results in some over-estimation of the inflation rate (Diewert and Fox, 2022).

1.1.2 The Paasche index

The *Paasche* index was introduced with the objective of solving the *substitution bias* present in the *Laspeyres* index formulation. Similarly to the latter, the *Paasche* index

takes the quotient of the weighted-average current price of the basket of components with the weighted-average base period price of said components. The difference, however, is that the *Paasche* index uses weights from the current time period in each weighted-average, effectively re-weighting the basket of components at each re-calculation of the index. Formally:

$$P_t = \frac{\sum_{c \in C} p_{c,t} * q_{c,t}}{\sum_{c \in C} p_{c,0} * q_{c,t}}$$

where:

P_t is the Paasche price index level in period t

C is the collection of goods and services in the price basket

$p_{c,t}$ is the price of component c in period t

$p_{c,0}$ is the price of component c in the base period

$q_{c,t}$ is the quantity, or weight, of component c in period t

With regards to addressing the most substantive limitation of the *Laspeyres* index, the *Paasche* formulation succeeds; the substitution effect is accurately accounted for, as the quantities, or weights, on each component update in every period of calculation. However, addressing this issue introduces two additional drawbacks.

Firstly, the *Paasche* index is far more data-intensive to compute. Given that current-period weights are used in each calculation period, this means that these weights must be observed and collected each time the index is due to update (e.g. on a monthly basis). In many cases, it may be difficult to obtain up-to-date consumption patterns on an equivalent frequency to the desired index computation schedule. This is the primary reason why the *Laspeyres* index remains a popular choice for key monthly publications in major nations, such as the consumer price index mentioned previously.

A secondary effect of this re-weighting methodology is that the *Paasche* index tends to do the converse of the *Laspeyres* index; it typically understates the rate of inflation. The constant re-weighting of the index results in decreases in product

consumption owing to price increases being immediately reflected by a drop in the product weights. As such, the components of the basket which would be the largest contributors to the inflation rate tend to be mechanically down-weighted before they do so, owing to the natural effects of market dynamics (Braun and Lein, 2021).

1.1.3 The Fisher index

The *Fisher* index is widely considered to be the optimal price index, as it manages to address the drawbacks of both the *Laspeyres* and *Paasche* indices simultaneously. It does this by simply taking the geometric mean of both of these indices. As a result, both of the methodologies are combined in an equally-contributory manner, reducing the opposing biases which they exhibit. Formally:

$$F_t = \sqrt{L_t * P_t}$$

where:

F_t is the Fisher price index level in period t

L_t is the Laspeyres price index level in period t

P_t is the Paasche price index level in period t

By using a geometric mean of the two formulations, the substitution bias is accurately accounted for, as the *Paasche*-portion of the formula uses the component weights from the current time period, which will account for consumers who switch to a cheaper competitor product following a price hike. Despite this, the largest contributors to the inflation rate will also not be entirely mechanically down-weighted, as they would be in the *Paasche* index, as the *Laspeyres*-portion of the formula retains the base period component weights (Braun and Lein, 2021).

Of course, the *Fisher* index comes with the inherent downside of requiring additional work to calculate, due to the fact that both the *Laspeyres* and *Paasche* indices must be computed in order to derive it. This also means that comprehensive, up-to-date consumption data is required in every time period in which the index is to

be recalculated, as the components must be re-weighted each time it is updated. For these reasons, the *Laspeyres* index generally remains the most popular choice for consumer price indices internationally, despite acknowledgement that a *Fisher*-form index would offer a superior measure of price change.

1.2 Challenges of building price indices for housing

One of the major challenges associated with property price indices, in comparison to other price indices, is that the stock which the index is attempting to measure is sold quite rarely and infrequently (Chandler and Disney, 2014). Only a small fraction of the total stock of of housing is sold in each calendar year, meaning that there are no guarantees of each year's set of sale transactions constituting an unbiased, representative sample of the entire housing market. The problem is exacerbated when considering sale transactions on a month-by-month basis, which is generally the minimum requirement of any property price index which aims to be useful to market stakeholders (Maguire, Miller, et al., 2016).

This raises a key question: which house price indices are attempting to measure the change in the average price of the housing market as a whole, as opposed to measuring the typical price of the specific set of properties transacted in a period of interest (Chandler and Disney, 2014). The former of the two concepts is a highly desirable metric to housing market stakeholders who wish to track the price of the asset class, while the latter is more akin to a measure of the mix of sales in each month, i.e., was a more expensive set of properties sold in month X than in month Y.

Many house price index models attempt to account for the issue of variations in the property mix by allowing the typical mix of properties over a period of time to dictate the weight for each type of property in the construction of the index for the current month (Silver, 2016). This is a somewhat similar approach to that taken by the *Laspeyres* index methodology, as discussed in Section 1.1.1. While this method is successful in controlling the impact of variations in the mix over the short term, the longer term effectiveness of this measure is not guaranteed; an arbitrary set of

transactions remains the determining factor in the index weighting, albeit one which is sampled over a longer period of time.

A phenomenon which could potentially impact the efficacy of this form of mix adjustment is the *starter home hypothesis*. This theory proposes that smaller homes with a lower value are likely to be sold more frequently than larger, more expensive properties. The reasoning given for this hypothesis is that the owners of these smaller homes tend to be younger and through career progression, become capable of moving up the property ladder over time (Costello and Watkins, 2002; Dorsey, Hu, et al., 2010; Jansen, Vries, et al., 2008; Ortalo-Magné and Rady, 2006). In contrast, those who have a larger, more valuable home tend to be more settled and satisfied and thus are less likely to sell the property. Data on property sale transactions over long periods of time supports the theory, demonstrating that these *starter homes* are an over-represented subset of the total number of sale transactions, relative to the number of said homes in existence (Clapp and Giaccotto, 1992; Costello and Watkins, 2002; Dorsey, Hu, et al., 2010; Jansen, Vries, et al., 2008).

As a result of the *starter home hypothesis*, any form of mix-adjustment based upon the number of sale transactions over a period of time (e.g. one-to-five years) is likely to suffer from bias towards these smaller homes, which are more affordable for first time buyers. This leads to under-representation of more valuable properties, which tend to remain off-market for very long periods of time.

Despite the critical importance of access to consistent, reliable indicators regarding the state of the property market, these factors, among others, have prevented experts agreeing on a consensus method of constructing a house price index and thus, a number of distinct methodologies are actively used by various institutions and national statistical agencies. Each of these methods comes with their own set of benefits and drawbacks, which will be discussed in more detail later in the thesis (Chandler and Disney, 2014; Maguire, Miller, et al., 2016).

The three most commonly studied and utilised methods of constructing a property price index are hedonic regressions, repeat sales regressions and stratified, mix-adjusted models (also known as central price tendency models) (Goh, Costello, and Schwann, 2012; Maguire, Miller, et al., 2016). Hedonic regression models and repeat sales regression models are classical methods which are very frequently used

by national statistical agencies around the globe, primarily due to their ability to access large amounts of rich property data as a result of their status as a government body (Hill, Scholz, et al., 2018; Scatigna, Szemere, and Tsatsaronis, 2014). Neither of these methods have undergone a great deal of evolution and improvement since their inception. Models which employ stratified, mix-adjusted methods, on the other hand, are newer and tend to be more frequently constructed by data-limited private entities who are producing their own property price indices, generally due to the typical untimeliness associated with national statistical offices' indices. Examples of popular stratified, mix-adjusted models in the United Kingdom include those from Acadata[1], Hometrack[2] and Rightmove[3].

Other actively used methods for measuring property prices exist, however, they are typically more simplistic measures which do not account for the variations in the mix of properties and thus are not attempting to measure the change in the total stock of properties, rendering them mostly irrelevant to the topic of research at hand.

1.3 Emerging research in the field

Given the importance of the property asset class to the economy and financial markets and the impact made by changes in the trend of said asset class, it is quite surprising that comparatively little research is undertaken on constructing new, superior modelling methods for the change in house prices. Rather, the majority of modern literature has focused on addressing the drawbacks and imperfections associated with the de facto standard hedonic regression and repeat sales models.

A variation on the repeat sales methodology which allows for the straightforward construction of regional sub-indices has recently been proposed (Larson and Contat, 2021). Previously, this was difficult to achieve, due to the data hungry nature of the repeat sales methodology, as discussed in Section 2.2. The nature of the methodology requires that all properties considered be sold at least twice in the sample period. Given that a relatively small number of properties will meet this criteria

[1] http://www.acadata.co.uk
[2] https://www.hometrack.com/uk/insight/uk-house-price-index/
[3] https://hub.rightmove.co.uk/latest-house-price-index/

on a month-to-month basis, generating sub-indices for a region generally results in too small a sample size for the production of an accurate, reliable model.

One of the major issues with the currently popular methods of measuring the change in property prices is timeliness. As stated previously, national statistical agencies often tend to publish their first estimate of monthly change one-to-two months after the fact. This makes it difficult for stakeholders to gain access to up-to-date information, thus hindering their ability to make accurate decisions regarding their stake in the market. It has been demonstrated that this untimeliness is unnecessary, as property listing data holds enough information about the market trends to deliver a highly correlated model to its corresponding repeat sales index, which is computed only on fully settled sales (Anenberg and Laufer, 2017). Such a change in the handling of data would allow for the repeat sales index to be computed with practically no lag.

An undesirable characteristic of hedonic regression models is the tendency for prior-month figures to revise substantially in future months. This is due to a number of factors, the greatest of which being the late-reporting of many property sale transactions, something which we have discussed with reference to the timeliness of model publications. According to recent literature, adjustments can be made to the initial model estimate based on analysis carried out on prior months, to reduce the impact of these revisions substantially (Sayag, Ben-hur, and Pfeffermann, 2022). This is a highly beneficial contribution, as the potential for large revisions compounded with the usual untimeliness of hedonic regression models makes it arduous to undertake important decisions based on the output of said models, leaving the stakeholder highly vulnerable.

A frequently ignored factor in house price modelling is the change in the quality or condition of dwellings over time. When houses are renovated in ways which do not show up in the typically selected regression variables (number of rooms, floor area, etc.), for example, by improving energy efficiency, installing solar panels, renovating rooms, replacing roofing, etc., the increase in value afforded by those changes may be misinterpreted by the hedonic regression model as genuine price growth. A study carried out by Reusens, Vastmans, and Damen, 2023 demonstrated that a novel way of incorporating the condition of a property into the hedonic regression

resulted in an improvement in the measurement accuracy of the price trend. Their analysis concluded that dwelling quality has been on an upward trend and the omission of this factor from models has resulted in overestimation of price growth.

Research into the data powering house price index models is another area of critical importance in achieving more accurate and timely measures of price trend. Detailed transaction data on housing is difficult to acquire, owing to property exchange still being a highly manual, administrative and protracted process. This is even moreso the case when it comes to new-build properties, which are regularly sold in advance of their construction being completed. Analysis by Hill, Pfeifer, et al., 2024 found that the inclusion of new-build properties in house price index models results in significant distortion of the index. As mentioned previously, the sale and, most importantly, sale price are frequently agreed upon for these new-builds whilst they are still being built. However, the transaction only enters the index calculation once the process has been legally completed and ownership has transferred to the buyer, resulting in heavily lagged prices entering the index calculation.

Despite these contributions, the classical implementations of hedonic regression and repeat sales models continue to be widely used by official, national indices, seemingly without much consideration for potential alternatives or improvements to said models and, in particular, their punctuality.

1.4 Modelling geospatial factors in housing

One avenue of research which has been somewhat neglected relative to the classical methodologies when it comes to house price indices is the field of geostatistics. Housing inherently exhibits a high degree of spatial auto-correlation, i.e. the prices of houses which are co-located tend to be highly correlated (Basu and Thibodeau, 1998; Cellmer, Cichulska, and Bełej, 2020). As such, properties located in the same locality tend to see their values move in tandem over time.

The cause of this behaviour is quite intuitive; the value of a house is highly dependant on environmental factors such as crime rates, the quality of local schools and public transport infrastructure, amenities and green-space and proximity to economic hubs, among others (Conway, Li, et al., 2010; Haurin and Brasington, 1996).

Therefore, we can conclude that location offers strong explanatory power over the value of a property.

Given that each of the aforementioned environmental factors would impact neighbouring properties in a very similar manner, it is evident then why spatial auto-correlation has been strongly observed in the housing market.

A common tactic employed in regression-based models which attempts to account for this spatial auto-correlation in house prices is the inclusion of dummy variables representing localities or neighbourhoods (Anselin and Lozano-Gracia, 2008). While these variables pass the statistical significance test in terms of their explanatory power, they are a blunt tool. For example, if the localities are defined to be very large, the ability for the variable to accurately capture local nuances will be diminished, as it is attempting to cover too many distinct environments. On the other hand, if the neighbourhoods are defined to be very small, there will not be enough samples on a month-to-month basis to achieve a sufficient goodness-of-fit, which will then introduce noise.

Furthermore, these dummy variables are not capable of applying a higher weighting to houses which are located closer than others in the locality. Assume two properties in the locality sell in a particular month: one property is at the furthest possible point away from a given house (without leaving the neighbourhood boundary), while the other is the neighbouring property. Using the system of dummy variables, both of these houses will have equal influence over the model estimate for the given house, despite the likelihood that the neighbouring property will give a superior estimate (Bala, Peeters, and Thomas, 2014).

As highlighted by Soltani, Zali, et al., 2023, property transaction data is, by nature, a nested and hierarchical dataset. Given that some property characteristics included in conventional hedonic regression models are specific to a given property (e.g. the floor area), while other characteristics apply to all dwellings within a particular region (e.g. crime rates, proximity to a good school, etc.), the attributes can be arranged in a hierarchy. Furthermore, co-located houses often share a common set of characteristics and therefore, property connections within a locality are nested. As their research concludes, these features of the dataset violate the independence

condition of a conventional regression model and thus the standard errors of the coefficients will be underestimated (Soltani, Zali, et al., 2023) under this methodology.

Some attention has been given to alternative methods of incorporating spatial factors into hedonic regressions. For example, a study by Cellmer, Cichulska, and Bełej, 2020 demonstrated that applying a method of adjusting the weights of observations in the hedonic regression fitting process based on their spatial relationships with other observations in the sample delivered a significantly higher goodness-of-fit measure than the classic hedonic regression model.

Moreover, their geographically-weighted hedonic regression model indicated that geospatial clusters where explanatory variables had the same or very similar values were highly prevalent in the dataset. This is not a surprising discovery, given the high degree of spatial autocorrelation present in housing, as discussed previously. However, this research indicates that many of the included explanatory variables may be unnecessary in order to achieve an accurate model, if this geospatial clustering behaviour was to be exploited more effectively and explicitly in the model construction.

Another variation of the geographically-weighted hedonic regression model which adds a clustering overlay was proposed by Verbic and Korenčan, 2017. In their research, they used a hierarchical clustering algorithm to group together areas in which localised hedonic regression models had demonstrated similar coefficients; indicating a certain degree of homogeneity. This allowed them to fit distinct hedonic regression models across more sensible geographical groupings, versus the common approach of using administrative regional boundaries. These boundaries are usually arbitrary and offer poor levels of explanatory power for neighbourhood quality in most hedonic regression models (Law, 2017).

The results of their clustering approach were a reduction in heteroskedasticity in the hedonic regression model and regional coefficients which are statistically-significantly different from the national model; indicating that the various characteristics of each property hold somewhat different levels of explanatory power on price in each of the homogeneous regional clusters. This is a plausible result; in a city, one might find that floor area comes at more of a price premium than in the countryside, therefore it would seem reasonable that the regression coefficient of

that attribute is larger in the city, for example.

Developing upon these approaches of enhancing hedonic regression models with geospatial data, recent research by Ding, Cen, et al., 2024 utilises a neural network in determining the geographical weighting for the hedonic regression model. Whereas Cellmer, Cichulska, and Bełej, 2020 used a simple Euclidean distance measure for determining the weights, this study offers a multitude of distance measures and allows a machine learning model to determine the optimal regression weighting. This methodological enhancement leads to further out-performance over the classic hedonic model specification.

Despite these geospatial enhancements, the majority of the studies discussed focus on applying an overlay to what is still a hedonic regression model at its core. Once again, we might question whether the full set of explanatory variables used in each hedonic regression model would be necessary if the geospatial effects were better integrated within the core model itself, rather than overlaid atop traditional methodology. Acquiring such a rich set of attribute data on each property transacted in each month is challenging, timely and costly. Alternative methods which do not rely on such an expansive dataset would likely be of great interest to many market stakeholders, as we will discuss in greater detail in Section 2.4.

1.5 Objectives of the research

The key objective of this thesis is to outline an alternative methodology for constructing a property price index model which does not rely on bolstering the shortcomings of a hedonic regression model. The proposed index is designed to leverage the inherent geospatial auto-correlation present in housing and will be built around a number of key facets which are intended to deliver a model which is more scaleable, automated, frugal and flexible than conventional methods. These critical features are outlined in List 1.

Prior to presenting the delivery of the research objectives themselves, Chapter 2 will introduce in greater detail the conventional methods used to produce house price indices, including their methodology, strengths and drawbacks. Furthermore,

LIST 1.1: Key *GeoPrice* index methodology features

- ☐ The model must be able to ingest a property transaction dataset and compute the house price index output without any human intervention, i.e. it must be fully automated.

- ☐ The model must be able to function with a bare minimum set of attributes; that is, the sale date, the sale price and a set of GPS co-ordinates for the property.

- ☐ The model should have the flexibility to operate on different types of property datasets, for example, using asking price listings rather than completed sales.

- ☐ The model should be capable of factoring in additional property attributes (e.g. number of bedrooms) to improve the accuracy if they are available to the user.

- ☐ The model should be scaleable and performant enough to be able to produce a high quality index on a region of any reasonable size, whether the number of monthly transactions is in the order of thousands, or hundreds of thousands.
  - ▷ This includes the ability to model sub-indices for partitions of the dataset, e.g. by region, or by additional attributes such as property type, where such attributes are available.

- ☐ The model must be capable of producing an index rapidly once the transaction data becomes available (i.e. within a day), to address the lengthy publication lag associated with conventional models.

- ☐ Despite these restrictions, the model must deliver equal or better performance in terms of noise reduction and smoothness than conventional hedonic regression models, even when using the minimal set of attributes laid out above.
  - ▷ The *GeoPrice* index should be highly correlated with the long term price trend measured by the benchmark, i.e. it must be evident that both indices are attempting to measure the value of the same asset class.
  - ▷ The metrics by which the *smoothness* of the *GeoPrice* index and the benchmark index are measured must be defined and justified.

the importance of monitoring house price trend for a variety of key market stakeholders including financial institutions, policy makers and asset owners will be explored, as well as the feasibility of each of these parties implementing one of the conventional methodologies to produce a bespoke house price index.

Following on from this, Chapter 3 will introduce the initial version of the *GeoPrice* model. Using a sparse and frugal transaction dataset of Irish property sales, the baseline index methodology will be outlined, in addition to benchmarking the results against the official national house price index of Ireland, which leverages a standard hedonic regression model. While this foundational version of the *GeoPrice* index succeeds in meeting the first two criteria outlined in List 1, it falls short on scalability, timeliness and performance, owing to inefficiency and slow execution time from the geospatial matching process.

The key focus of Chapter 4 will be to investigate methods of addressing these efficiency concerns. The *GeoTree* is a bespoke tree-like data structure which stores property sale transactions in nested, hierarchical clusters. At each level of the tree, properties are broken up into buckets based on their proximity to one another. As one traverses deeper down the tree, the size of the buckets become progressively smaller, allowing the user to hone in on groupings of proximate houses which can be used as points of comparison when leveraging the spatial auto-correlation effect of neighbouring properties.

By introducing caching at each node of the tree, a collection of pointers to every leaf node found within the sub-tree beneath that node allows for a rapid, $O(1)$ enumeration of a collection of neighbouring properties within a certain distance bound of any transaction in the dataset. The introduction of this data-structure gives the *GeoPrice* model the technical ability to fulfill the scalability and timeliness goals set out in List 1, with the execution time improving by multiple orders of magnitude; completing the *GeoPrice* index calculation in a matter of minutes, rather than days.

With the efficiency bottleneck removed, Chapter 5 will add additional flexibility through methodological enhancements. In addition to demonstrating the ability for the model to compute an index on a bespoke dataset of asking prices from a property listing platform, this chapter will augment the *GeoPrice* model with the ability to incorporate data on the number of bedrooms in each property.

Furthermore, a number of metrics by which the *smoothness* of the indices can be measured and compared against one another will be outlined. These smoothness metrics demonstrate that the combination of number of bedrooms with geospatial matching significantly outperforms the geospatial-only index. The methodological enhancements introduced alongside the use of an asking price dataset demonstrates the capability of the *GeoPrice* model to meet the flexibility goals outlined in List 1.

Chapter 6 will expand the *GeoPrice* model to a much larger region: the United Kingdom. Through the use of a publicly available dataset of sale transactions, this chapter will illustrate the ability of the model to produce highly correlated results to a hedonic regression model fit on the same transactions, without the need for the additional, rich attribute data on each property used by said model. Furthermore, it will also demonstrate that the *GeoPrice* model can achieve significantly greater levels of smoothness and a less noisy house price index than the benchmark hedonic regression model.

The ability of the index to outperform conventional models in distinct regions with different property dynamics, transaction volumes and market behaviours demonstrates its flexibility and scalability. Chapter 6 will also present regional sub-indices for each administrative region of England, in addition to sub-indices for each distinct property type, thus achieving the final remaining key feature of those set out in List 1.

Finally, the thesis will conclude with Chapter 7, which will discuss the delivery of the model objectives set out in List 1 alongside future avenues of research and potential applications of the thesis findings.

1.6 Chapter summary

Price indices are a critical tool for measuring a wide range of price trends in goods and services, which feed into crucial econometric data releases such as consumer price indices, gross domestic product and a wide range of other inflation-adjusted metrics. A variety of methodological formulations of price indices exist, each with their own distinct drawbacks.

One important area of application of price indices with particularly unique modelling characteristics is that of property price indices. Unlike most goods and services, such as food, clothing, medical expenses, etc. where taking a monthly sample of the cost is feasible, individual properties typically transact infrequently and their value cannot be straightforwardly measured outside of these transactions. Given that each house is unique, the set of houses sold in month X cannot be directly compared to those sold in month $X-1$ or $X+1$, leading to modelling challenges.

This lack of comparability in property transaction samples has given rise to a number of commonly used, specialised techniques for measuring house price growth, which will be discussed in detail in the coming chapter. Furthermore, the importance of these models to various real-estate stakeholders will be analysed, along with the viability of each of these parties implementing said models.

Chapter 2

Background: Property price index methodologies

As discussed in Chapter 1, there are a number of unique challenges associated with modelling house price indices owing to their distinct market dynamics. In this chapter, the most commonly used methods of computing house price indices will be outlined, along with each of their benefits and weaknesses.

Following on from this, the significance of accurate models of property prices to various market stakeholders will be discussed, in addition to the feasibility of each of these parties implementing the conventional methodologies discussed in this chapter.

2.1 Hedonic regression models

Hedonic regression is a method for constructing a property price index whereby each house is broken up into a large number of constituent characteristics (e.g. number of bedrooms, bathrooms, floor area etc.), each of which is given a weight according to how much impact said attribute has on the sale price of the property (Kain and Quigley, 1970; OECD, Eurostat, et al., 2013). These attribute weights are determined by fitting a regression using the sale transactions, with the characteristics of each property serving as either categorical or continuous exogenous variables in the regression, where appropriate.

While the core methodology is similar within hedonic regression models, a number of variations in the method of use of the model for measuring house prices exist, the most common of which being: the characteristics approach, the imputation approach and the time dummy regression approach (OECD, Eurostat, et al., 2013; Silver, 2016).

Both the characteristics approach and the imputation approach fit separate regressions across a range of successive time periods, typically monthly, with each of these regressions being fit on the property sales transacted in the corresponding time period. We denote these regressions as $R_0, R_1, \ldots, R_t$, for time periods $0, \ldots, t$. Once R_x has been fit, the coefficient attributed to each characteristic by the regression signifies the monetary amount which said characteristic contributes to the value of a property, if it were to be sold in period x.

Suppose, for ease of example, that each of these regressions are fit using only the number of bedrooms and square footage of each property as attributes. In reality, a much larger set of attributes is typically used. Under the characteristics approach, the average value of these attributes in time period 0 would then be determined for the region in question; let's assume that the average number of bedrooms is 3.2 and the average floor area is $1,473$ square feet. It is then possible to estimate the value which would be attached to said typical property if it were to be sold in time periods $0, \ldots, t$ by using the coefficients determined by regressions $R_0, \ldots, R_t$. The difference in the price level determined for each time period can then be used to generate a property price index.

The imputation approach operates on a lower level than the characteristics approach, looking at individual properties, rather than a typical property. Where the characteristics approach answers the question of *what would a typical property be valued at if sold in periods* 0 *through* t, the imputation approach addresses the question of *what would a particular property (with its constituent attributes) sold in period* 0 *be worth if revalued in periods* 1 *through* t (Silver, 2016). In this system, a fixed basket of properties is revalued in each time period using $R_0, \ldots, R_t$, with the average of the imputed prices of the basket being compared across time periods to generate the index. By adjusting the fixed basket of properties used, this approach accounts for changes in the mix of properties in a given month, such that an identical mix is compared across

time periods.

The time dummy hedonic regression model differs from the other two approaches such that only a single regression is used, rather than individual monthly regressions. Each time period $0, \ldots, t$ is given a dummy variable in the regression; effectively capturing the temporal component of the price within a set of coefficients. The variable takes on a value of 1 where the property in question was sold in the time period corresponding to the variable, otherwise taking a value of 0. The result of this is that the regression coefficients attributed to each of the attributes of the properties in the sample are fixed across time periods $0, \ldots, t$, instead of being allowed to vary as in the characteristics and imputation methods. Instead, the change in property prices between time periods is implicitly captured in the coefficients attached to the time dummy variables in the regression, as below (OECD, Eurostat, et al., 2013).

$$HPI_0^s = \exp\left(\hat{\delta}^s\right)$$

where $\hat{\delta}^s$ is the regression coefficient attached to the time dummy variable for the time period s.

While simpler to implement, use of the time dummy variable method is, in a sense, a double edged sword. On one hand, the fact that a single regression is used for the entire time range leads to a high degree of stability and robustness, as the regression has been fit on a much larger pool of property transactions (OECD, Eurostat, et al., 2013). The result is a model which has been able to very precisely determine the contribution of the various property characteristics to the sale price. On the other hand, one could argue that it is unrealistic to assume that the characteristics will retain the same weight over long periods of time, which is an implicit assumption of the time dummy variable method (Silver, 2016). However, this issue can be worked around by using techniques such as chained rolling time windows for the regression, where the oldest time period drops out of the model each time a new time period is added. As such, the size of the time period over which the regression ranges is fixed and thus the assumption becomes one of the attributes retaining the same weight over said fixed period of time, rather than over the entire history of the index.

2.1.1 Strengths and drawbacks

Hedonic regression models are very popular due to a wide array of strengths. Accuracy is a key concern when it comes to constructing a property price index. Hedonic regression models typically achieve an R^2 goodness-of-fit score in the range of eighty to ninety percent, assuming a sufficiently rich set of regressors are specified for each property. This leads to a property price index which is consistent and reliable.

Another appealing attribute of hedonic regression models is their ability to generate imputed prices for combinations of characteristics which may not necessarily trade during the period. For example, in cases where you have a region with a relatively small number of properties trading in each time period, it may be the case that you have certain combinations of attributes which do not appear in the sale transactions for every time period, e.g. there may not be a *4 bedroom, 3 bathroom detached property in a particular locality* sold in each and every time period. Despite this, the hedonic regression model can generate an imputed price for this type of property in every time period, by drawing on the knowledge it has acquired from other properties with a subset of those characteristics (OECD, Eurostat, et al., 2013; Silver, 2016). This ensures that a complete set of imputed prices can always be generated, regardless of the input sample, assuming that every potential attribute value has at least one sale containing said value.

Finally, the ease of creating sub-indices without the need to fit a separate model is a major draw to hedonic regression models, particularly the imputation approach. By altering the fixed basket of properties to include only those in a certain region, for example, one can generate a property price sub-index for said region. The same principle can be applied to other attributes to generate any type of sub-index which is desirable, e.g. an apartment index, a three-bedroom index, a new-build index etc. (Gouriéroux and Laferrère, 2009).

Despite its strengths, the hedonic regression model also comes with some significant drawbacks, which makes it infeasible to use for many institutions and market stakeholders. One of the most significant of these is the labour and experience required to apply the model. Due to the great deal of mathematical and computational complexity in the methodology, hedonic regression models are usually produced

by a team of full-time, experienced statisticians (Haan and Diewert, 2011; Maguire, Miller, et al., 2016). It is highly important that outlier and erroneous values are removed from the dataset in order to avoid contamination of the hedonic regression model and a number of concerns regarding normality of the dataset must be addressed and mathematically corrected for if not present.

Another major issue with hedonic regression indices is the richness of data required to build an accurate model (Larson and Contat, 2021). At a minimum, the regression would generally require data on the *new build status, property type, number of rooms, geographical locality* and *floor area* (OECD, Eurostat, et al., 2013). However, this set of explanatory variables could potentially expand to dozens of attributes including information about garden size, neighbourhood quality, local educational facility quality, public transport availability and any other data point which could have a significant impact on the sale price (He, Wang, et al., 2010). For example, according to Luttik, 2000, the presence of an attractive landscape next to a house was responsible for a premium of 5-12% over a property located in a less pleasing environment, with the other characteristics of the homes being otherwise the same.

This makes specification of a complete set of regressors extremely challenging and the large number of free parameters available to tune in the model can lead to over-fitting (Maguire, Miller, et al., 2016). This requirement for rich attribute data for each and every property in the sample makes the model impractical to use for most parties aside from national statistical agencies, as such data is typically not publicly available. The solution employed in many applications of hedonic regression models is simply to ignore this issue, which likely results in distortion and noise among the model coefficients.

A lack of timeliness is also a common issue with indices backed by hedonic regression models. National statistical agencies typically publish these models with a lag between one and two months, leading to information which is potentially out of date by the time it is released (Maguire, Miller, et al., 2016). The main reason for this lag is due to the complexity of the model, with the data requirements alongside the complex methodology demanding a significant time investment before results are available (Haan and Diewert, 2011). Property stakeholders who require up to

date information on market movements thus may not be satisfied with the timeliness of hedonic regression models and may need to look elsewhere to obtain up to date guidance on the property markets.

Furthermore, a key flaw to consider in using hedonic regression models is the inherent spatial autocorrelation of properties. Homes are generally considered to have some degree of spatial correlation, owing to the fact that common neighbourhood characteristics, such as crime rate, quality of education, public transport links and green-space, can have a substantial impact on the value of a given property (Conway, Li, et al., 2010; Ismail, 2006). These factors are frequently ignored by hedonic regression models, as they are too difficult to model and observe from month-to-month. However, disregarding this idiosyncrasy of the housing market may violate one of the fundamental assumptions of a linear regression; that the residuals should be independent of one-another (Cartern and Haloupek, 2000).

2.2 Repeat sales models

Repeat sales regression models measure the change in property price over long periods of time by using repeated sales of precisely the same property. Through this methodology, the model no longer needs to consider any inherent or external price-affecting attributes of the properties on which it is being fit, as matching only identical properties ensures that a like-for-like comparison is being made (OECD, Eurostat, et al., 2013).

In a sense, repeat sales represents the most authentic analogue of the matched model method of measuring inflation in more liquid products, such as food, clothing or car prices. As a result, only the sale date, address and sale price of each property is required in order to use the repeat sales method on a dataset of property sale transactions. Suppose that p is a property which has been sold multiple times in the time period $0, \ldots, t$, over which the model is covering; call two such time periods x and y, where $x <_{time} y$. Then we can construct a repeat sales model where:

$$\log\left(\frac{p^y}{p^x}\right) = \sum_{i=0}^{t} \beta^i \mathbb{D}^i(x,y) + \epsilon^i$$

In this case, the natural logarithm of the price ratio of property p in period y to period x is equal to a linear combination of dummy variables $\mathbb{D}^i$, defined as:

$$\mathbb{D}^i(x,y) = \begin{cases} 1, & \text{if } i = y \\ -1, & \text{if } i = x \\ 0, & \text{otherwise} \end{cases} \tag{2.1}$$

In other words, the dummy variables are given a value of 1 in the variable representing the current sale period, y, and a value of -1 in the previous sale period, x. The ϵ^i are error terms. In practice, a linear regression fit on the property sale transactions is used to estimate the dummy coefficients, thus replacing the β^i coefficients with estimates $\hat{\beta}^i$, (OECD, Eurostat, et al., 2013).

Once the $\hat{\beta}^i$ have been estimated for periods $0, \ldots, t$, the change in property prices from month 0 to month s can be calculated using the exponent of the regression coefficient corresponding to month s:

$$HPI_0^s = \exp\left(\hat{\beta}^s\right)$$

2.2.1 Strengths and drawbacks

Undoubtedly the greatest strength of the repeat sales model is the simplicity it offers. The model is conceptually very easy to understand; essentially measuring the change in price of the same property over differing time periods. Difficult problems which arise in other property price index models, such as controlling for variations in the mix of properties over different time periods, need not be considered in the repeat sales model (OECD, Eurostat, et al., 2013).

This methodology also requires a very frugal amount of information for each individual property, unlike the hedonic regression model described in Section 2.1. This frugality allows repeat sales models to be fit on publicly released property sale data in many countries, making the model more accessible to entities without access to the confidential information often afforded to national statistical agencies (Larson and Contat, 2021).

Furthermore, the use of a more conservative amount of data on each property reduces the risk of the model being overfit. A repeat sales model does not need to concern itself with the impact of green-space or attractive landscape not being incorporated into the model, as it only analyses the price of the same property, sold in multiple time periods. Granted, it is possible for these factors to change over time, however, this is a considerably less prevalent tail risk than simply disregarding the issue entirely, as hedonic regression models are often constrained to doing (Case, Pollakowski, and Wachter, 1991).

Unfortunately, the repeat sales methodology also brings a number of major drawbacks to the table. While there is a great deal of frugality in terms of attribute data required for each property, the model requires a vast number of property transactions to achieve accurate results, as properties which have not been sold more than once must be disregarded. In fact, in many instances, as much as 96% of the dataset may need to be discarded due to incompatibility with the repeat sales methodology (Case, Pollakowski, and Wachter, 1991). While this is not an insurmountable issue for a country the size of the United States of America, for example, it may leave smaller countries unable to utilise the method due to not having a sufficient number of repeated sales in each month of their dataset (Larson and Contat, 2021).

It has also been theorised that the sample of repeat sales is not representative of the housing market as a whole. For example, in a study by (Jansen, Vries, et al., 2008), only 7% of detached homes were resold in the study period, while 30% of apartments had multiple sales in the same dataset. This is argued to be due to the *starter home hypothesis*, previously discussed in Section 1.2 (Costello and Watkins, 2002; Dorsey, Hu, et al., 2010; Jansen, Vries, et al., 2008). This leads to over-representation of inexpensive and poorer quality properties in the repeat sales method. Cheap houses are also sometimes purchased for renovation or are sold quickly if the homeowner becomes unsatisfied with them, which contributes to this selection bias (Jansen, Vries, et al., 2008). Furthermore, newly constructed houses are severely under-represented in the repeat-sales model as a brand new property cannot be a repeat sale unless it is immediately sold on to a second buyer (Costello and Watkins, 2002).

A further major issue with the repeat sales method is the inability to account for depreciation of a property, nor renovations which occur between sales (McMillen

and Thorsnes, 2006). It is possible that a house undergoes extensive renovations after a sale, before eventually being resold at a profit. In the repeat sales model, no distinction will be made between such a scenario, and a standard sale without a renovation taking place. As such, the increase in value imparted by the renovation will be considered to be a genuine appreciation in house prices, rather than value being added to the property in question.

2.3 Mix-adjusted models

Models which employ mix-adjustment and stratification techniques are often simpler and more accessible to implement for private entities, particularly those with limited data. This class of methods spans a wide range of model complexity, stretching from the most basic indices constructed using a simple median of sales transactions, to more complex models using detailed stratification methods in an attempt to adjust for bias in the quality mix between time periods (Maguire, Miller, et al., 2016; OECD, Eurostat, et al., 2013).

These models often rely on the law of large numbers, i.e., large sets of clustered data often exhibit a noise-cancellation effect, where errors become small after aggregation (Maguire, Miller, et al., 2016). As a result, a tolerance for approximation in comparison of properties can be exploited to allow the use of the entire dataset, unlike repeat sales, without requiring a rich suite of attribute data for each and every property in the dataset, as needed in hedonic regression.

A commonly employed strategy for reducing the sample bias between time periods is a basic mix-adjustment. While a detailed stratification usually cannot be carried out due to a lack of attribute data, location-based stratification alone offers substantial benefits to the reliability of a mix-adjusted model, as will be demonstrated in this thesis.

Suppose the dataset is split into buckets according to the town in which the property fall; call these buckets B_s^r for $r \in R$, the set of all towns, with properties from time period s. We can construct a stratified mix-adjustment index using these strata as follows:

$$HPI_0^s = \sum_{r \in R} \frac{\mu_{B_s^r}}{\mu_{B_0^r}}$$

where $\mu_{B_x^r}$ is the mean, median or any other desired aggregation method of the bucket of properties in region r, sold in time period x. As such, these stratification method compares properties on a region-by-region basis, before aggregating the result.

This stratification method can be improved by adding weights according to the prevalence of property sales for the region in question (Wood, 2005). If a lengthy history of data is available to the model, weights can be computed for region r by calculating the ratio of property sales in region r to the total number of property sales across the same period of time. As such, we can define the regional weights:

$$\omega_r = \frac{\sum_{i=0}^{t} |B_i^r|}{\sum_{x \in R} \sum_{i=0}^{t} |B_i^x|}$$

where $|B_i^x|$ is the size of the bucket of sale transactions for region x in time period i. Using these geographically stratified weights, we can redefine the stratified property price index model as:

$$HPI_0^s = \sum_{r \in R} \omega_r \frac{\mu_{B_s^r}}{\mu_{B_0^r}}$$

If more attributes are available to the model, such as property type, new build status or year of build, these can be further incorporated to the stratification model, with weights adjusted appropriately for each bucket (Wood, 2005).

2.3.1 Strengths and drawbacks

The mix-adjustment method's greatest strength lies in the flexibility and adaptability of the methodology. Where the hedonic regression model requires a great deal of rich property data to deliver reliable results, mix-adjustment models with stratification can be tailored to make the best of any level of data frugality. The method also makes use of all transactions present in the sale data, unlike repeat sales, which must disregard a large amount of the dataset. This makes the mix-adjustment method

more suitable for smaller datasets or datasets which stretch over a relatively narrow period of time (Prasad and Richards, 2008). Furthermore, this ability to operate on the entire sample of transactions inherently reduces the risks of bias and overfitting of the model (Babyak, 2004).

The simplicity of the concept is also a strong draw to mix-adjusted methods, similar to the repeat sales method. Unlike hedonic regression, it is conceptually straightforward to understand and implement a basic mix-adjusted stratification model on a rudimentary, publicly available dataset. The proficiency and resources available to the user of the model can be used to add additional complexity, or to improve the performance and stability of the model accordingly. This increases the accessibility of the model to a broad set of use cases, rather than being restricted to those with expertise in specific technical subject matters (Prasad and Richards, 2008).

Similarly to hedonic regression models, it is easy to produce sub-indices for any of the strata in a mix-adjusted model. By simply excluding the strata which do not match the desired sub-index criteria, one can generate a house price sub-index for any desired subset of the property data, for example, an index for apartments or detached houses. As discussed previously, the concept of mix-adjusted median house price index models can be thought of as constructing sub-indices on a stratum-by-stratum basis and adjusting the mix by altering the weighting of each stratum such that the composition of strata remains comparable over time (Haan and Diewert, 2011).

While the method is less demanding than hedonic regression when it comes to the need for rich attribute data for properties in the same, mix-adjusted and stratified methods have tended to require a certain amount of data on each property, in order to adequately separate properties into appropriate strata (OECD, Eurostat, et al., 2013). This is in contrast to the repeat sales method, which requires no attribute data and requires only the address, date of sale and sale price, yet over a considerably longer window of time.

Another issue with this method is that stratification can result in very small samples per strata, if number of attributes used in the grouping is extensive. This could potentially result in unrepresentative strata, due to an insufficient number of samples to generate an accurate average price for the group in question (Turner, 2003).

This would pose less of an issue in cases where the total pool of properties in the market is very large. However, for a smaller country like Ireland, this is an issue which not only would plague a mix-adjusted median approach, but would hinder any house price index model (O'Hanlon, 2011).

Similarly to the repeat sales method, stratification usually does not account for depreciation or renovation of properties, unless particular attention is given to handling this issue. This results in a similar scenario to that of repeat sales, where an increase in value due to the reconditioning of a property will be considered by the model as genuine property market inflation. This factor is highly difficult to account for and is usually disregarded even in hedonic regression models (e.g. the Office for National Statistics UK's house price index) where it would theoretically be possible to account for (Anderson, 2018).

2.4 Importance to real-estate stakeholders

Property price indices are understandably of great interest to financial institutions, particularly banks who partake in mortgage lending and thus are exposed to a great deal of asset risk through the collateral on these loans (Miller and Maguire, 2022).

However, property market trends are also closely watched by many other parties, including home-owners, home-buyers, businesses, central banks and governments. In this section, we will outline the importance of these models to each of these stakeholders, as well as exploring their feasibility of implementing each of the models discussed in this chapter.

2.4.1 Financial institutions

While there are a multitude of stakeholders in the property market, perhaps the greatest of these is the financial services sector, due to the risk taken on through mortgage lending. For the majority of people, a house is the most valuable asset they will own in their lifetime. Furthermore, almost one-third of British households are actively paying a mortgage on their house, which collectively forms the greatest source of debt for said group of people (Bank of England, 2018a).

Mortgages are a key source of revenue for banks and financial bodies, due to their long repayment length, which results in a considerable amount of interest accrued. However, they also pose a substantial risk for said financial institutions, as they involve the lending of a large principal which is often repaid over decades, during which the financial circumstances and stability of the borrower are not guaranteed to remain constant and indeed, are often influenced by the flux in property prices as an indicator of general economic stability. This makes it difficult to predict the number of borrowers who will struggle to meet their repayments during periods of economic downturn (Bank of England, 2018b).

While an economic recession usually results in massive downward pressure on commercial property prices and the equities market, such a sharp drop tends not to be reflected *quite* as drastically in the residential property market. Rather, the number of sale transactions usually drops, as property owners no longer wish to sell their house for a lower sum of money than they would have received before and many will delay their decision to sell. It is likely that such a drop in residential property sales volume is reflected in a reduction in new mortgage applications, hence resulting in a loss of revenue and profit for lenders. Furthermore, such an economic event signals reduced financial stability for borrowers and thus default rates on mortgages will rise, causing a greater amount of bad debt on the books (Bank of England, 2018b; Zhu, 2005).

It is logical then that financial bodies are highly interested in tracking the movements in property prices, to inform their lending policies and risk assessment methods. A more bullish property market may lead to banks taking on slightly more risk, with a view that the property will appreciate and so too will the confidence of the borrower. Conversely, a bearish market will likely result in a tightening of the lending criteria, with institutions only taking on highly financially secure borrowers who they judge to be capable of weathering the storm of further depreciation of their newly-purchased property, in a worst-case scenario (Che, Li, et al., 2011; Guerrieri and Uhlig, 2016). They might also be interested in comparing a mortgage application to the average property price for that region, to judge whether the price is excessively expensive when balanced with the financial circumstances of the applicant.

The untimely manner in which government statistical offices tend to release information on market movements, with a lag of one to two months being typical, may result in key policy decisions around lending being made later than is ideal. As a result, larger financial institutions are often interested in creating their own custom house price model which delivers up-to-date information, in order to better inform their lending criteria (Miller and Maguire, 2022). We will present a suitable, performant model meeting these criteria later in this thesis.

2.4.1.1 Modelling viability

Where a bank wishes to develop their own property price index model in order to get more up-to-date market information, there are some key considerations when choosing the appropriate methodology to employ. While the repeat sales method might at first seem tempting due to the simplicity of implementation, further thought reveals that this method is unlikely to be suitable. This algorithm relies on comparing multiple sales of the exact same house over long periods of time. If a financial body is using their historical mortgage data to fit the model, it is unlikely that the past sales of any given property were conducted using mortgages taken out at the same bank by different buyers, resulting in a low match rate for what is already a wasteful method in terms of data utilisation. Furthermore, historical data stretching back over decades is generally necessary to generate a reliable result with this method, which will likely be difficult for an institution to both source and convert into a clean, rich digital format (De Vries, de Haan, et al., 2009).

The hedonic regression model may be a viable option, as these institutions will have property characteristic data for the properties on their loan books, which is key to the performance of this algorithm. However, the main drawback of using this method is the complexity of the model. The process of creating a hedonic regression model is very theoretically intense and generally requires the work of a number of statisticians in order to implement and interpret the index on an ongoing, regular basis. Furthermore, due to the human labour associated with maintaining a hedonic regression model, as well as the reliance on rich, detailed and well filtered data, it is difficult to produce the model on a more frequent time schedule than monthly

or bi-monthly, particularly when this work must be repeated on a region-by-region basis, where an institution wants more granular measures than a national model.

Overfitting is another possible avenue of concern with regard to hedonic regression indices, as mentioned in Section 2.1. As hedonic regression relies on having a complete view of the property market, it may adapt poorly to financial institutions who likely only have access to a biased sample of property sales which have used their own lending products as the method of payment. If a particular bank was to target the middle-class working family as their intended customer base, for example, this may lead to a bias in the type of homes which are predominantly included in the model's data pipeline, thus not accurately capturing the trend in the broader housing market, rather, only the movements in a subset of it.

Mix-adjusted median based property price index models may therefore prove the most effective option for a financial institution to implement. The main advantages of such an approach lie in the ease of implementation and flexibility to incorporate various data sources of differing densities. Firstly, a mix-adjusted median algorithm can usually be computed in an entirely automated way, without a great amount of tuning or manual processing, reducing the need for multiple statisticians to spend time constantly tweaking the model to produce a monthly release, particularly where results are being produced for a number of different cities or regions. This allows for the model to be recomputed very frequently; as often as daily or two-to-three times per week, if sufficient *live* incoming data is available for the model.

This model also does not rely on specifying a complete set of price-affecting characteristics and can work with as little as three attributes: the sale date, the address and the price. Due to this, the algorithm can use the entire property sale transaction data for greater accuracy and avoidance of overfitting, which is published publicly in most countries; for example, by the Property Services Regulatory Authority in Ireland (Property Services Regulatory Authority (IE), 2024), or by HM Land Registry in the United Kingdom (HM Land Registry, 2024). Furthermore, the flexibility of the methodology allows for additional core attributes, such as the number of rooms, to be included for greater accuracy, as we will demonstrate later in this thesis. This means that the institution can mix their own highly detailed mortgage data

together with general, unbiased but sparsely-detailed data for property sales, in order to increase the model's perspective of the market as a whole. As a result, the mix-adjusted median model is a sensible option for large banking institutions who wish to see very regular updates on the market in order to aid them in deciding on their lending policies.

2.4.2 Homeowners and homebuyers

On the opposite side of the spectrum of stakeholders, you have active homeowners and prospective homebuyers. Despite often being on opposing sides of a financial contract with the financial institutions discussed previously, homeowners share much in common with their counter-parties in terms of the impact of moves in the property market. An increase in the value of property prices is a positive outcome for active homeowners, as their investment in property generates unrealised gains. This makes the likelihood of defaulting lower, as it results in a lower incidence rate of *negative equity* (i.e. the present market value of the home falling below the amount owed to the lender on said home), which is correlated with mortgage defaults (Elul, Souleles, et al., 2010; Labonte, 2007). This is a benefit to the lender, as they experience less risk on their lending activity.

Furthermore, it is demonstrated that strong upward moves in the value of properties are correlated with increases in the incidence of turnover in the market, i.e. there is a higher probability of existing homeowners looking to sell their home and crystallise the gains associated with the investment according to the magnitude of their unrealised returns (Tu, Ong, and Han, 2009). Thus, it stands to reason that homeowners are interested in monitoring the state of the property market in real-time, in order to estimate their return on investment and to judge their appetite to sell.

Property prices are long understood to experience *boom-bust cycles*, i.e. a period of strong increase in asset values, followed by an oftentimes protracted period of price correction (Labonte, 2007; Nofsinger, 2012). A more timely property price index model would result in the ability for owners and prospective buyers to better speculate on the transition from *boom* to *bust* and attempt to time their sale or purchase, respectively. This kind of speculation is commonplace in every other asset

class traded on the financial markets, however, the lack of accessible data and tools makes it significantly more challenging in the property market. Conversely, homebuyers may wish to attempt to time the market, in order to avoid purchasing near the peak of a boom cycle and landing in negative equity, should they take out a mortgage and subsequently experience a fall in the value of their newly acquired asset.

2.4.2.1 Modelling viability

The implementation of any method of producing a house price index is generally not feasible for the typical homeowner or prospective buyer. At present, these stakeholders are typically at the mercy of the indices produced by the national statistical offices of their respective country. As discussed prior, these statistics are published with a lengthy publication lag and are difficult to verify for the layperson, due to the complexity of the methodology and lack of availability of the data feeding the models (Miller and Maguire, 2022).

On the other hand, a more simplistic mix-adjusted median property price index, such as the one being proposed in this thesis, is theoretically reproducible by any homeowner, due to the ability of the model to work solely with publicly available datasets. While there are some limitations in terms of the technical complexity of implementing the indices (e.g. via a programming language), this is the only significantly restrictive factor associated with use of this type of model.

The primary application of this research with respect to homeowners and homebuyers is the opportunity for more providers to create their own property price index, which may either be public facing and directly available to said stakeholders, or integrated into a system which directly or indirectly benefits them. For example, more property portals such as *MyHome* [1], *Zoopla* [2] and *Daft* [3] may become capable of implementing their own market index, using their in-house data, which could give their customers a better indicator of the estimated value of listed properties, or their own property, when buying or selling.

[1] https://www.myhome.ie/
[2] https://www.zoopla.co.uk/
[3] https://www.daft.ie/

2.4.3 Businesses and the property development industry

Another key stakeholder in the property market lies in commercial business, for a variety of reasons:

- A business may be considering the purchase of property as a business asset.
- A business may be considering the disposal of a property asset to increase liquidity or target an alternative investment.
- A business may be operating in an industry directly impacted by property price fluctuations, for example, property developers.

For many businesses, property is the largest asset (by value) on the balance sheet (Liow and Ooi, 2004). This is even more apparent for property developers, whose *trading stock* is also in the form of housing. It would stand to reason then that fluctuations in the market have a significant impact on the majority of businesses and it is important for said businesses to attempt to forecast and plan around these market moves.

Property cycles often drive business cycles, with a sharp decline in asset prices typically coinciding with an economic recession (Labonte, 2007). Given that recessions result in a decrease in cash flow and a restriction of resources for the majority of businesses, the management of property on the balance sheet is a key aspect of recession preparation and business survival (Goldberg, Phillips, and Williams, 2009). Firms, particularly those who may have invested in property using excess cash during a boom cycle, may wish to sell and raise additional cash reserves to aid any *weathering of the storm* which may be required in an uncertain economic environment (II and Michael, 2006). If businesses wait until they run into cashflow problems, it is likely that they will be forced to sell at a substantially lower value (potentially at a loss), with a lower number of prospective buyers available delaying the completion of the sale (De Wit, Englund, and Francke, 2013).

For property developers, the market becomes an even more crucial aspect of business to consider. While for the typical business, the concern around property is primarily focused on the value of their existing assets, for a property development firm, they are faced with a situation where:

- Their turnover will reduce substantially due to a lower incidence of purchases during a property bust cycle (De Wit, Englund, and Francke, 2013).
- The final sale value of their active projects is likely to be considerably lower than their initial projections, despite most of the costs already having been fixed (Labonte, 2007).
- Future projects are likely to be cancelled and/or scaled back in order to protect the business during the downturn (Jones and Evans, 2013).
- Developers who lease some of their property portfolio will experience a drop in rent rates (Grover and Grover, 2013).
- Their assets on the balance sheet, be it commercial property or private homes, will lose value (Grover and Grover, 2013).

Developers may wish to attempt to forecast the market in order to decide upon projects to greenlight and the size of their active workforce, in addition to being able to produce more accurate and up-to-date prospective sale values estimates when performing profitability evaluations on project proposals.

2.4.3.1 Modelling viability

Neither the hedonic regression nor the repeat sales property price index models would be a viable undertaking to implement for the vast majority of property developers. A bespoke hedonic regression model would be difficult to support without a number of full-time statisticians and software engineers, which property developers are unlikely to be capable of sustaining among their staff. Furthermore, hedonic regression models require a detailed and varied dataset in order to achieve a good fit, whereas the data available to developers is likely to be too narrow and similar, as it will consist of new build property concentrated in a few key areas (Maguire, Miller, et al., 2016).

The repeat-sales method, on the other hand, requires a considerably long history of data which most developers are unlikely to have, given that most of those that existed prior to the Global Financial Crisis went bankrupt (Jones, Cowe, and Trevillion, 2018). However, even for those who are long established, they would only have

data on the initial sale of each property, whereas the repeat sales method measures the change in value of the same property, re-sold in different time periods. This rules out the use of this model for property developers.

The mix-adjusted median model could be used on a combination of in-house data and publicly available property sale records in order to produce a price index. While this would likely require a software engineer, the human resources needed would be considerably less than what would be required to fit and maintain a hedonic regression model.

2.4.4 Governments and regulatory bodies

A change in the trend of house prices can have an extraordinary impact on the general strength or weakness of an economy. When property prices are high, homeowners feel secure in increasing both spending and borrowing, which in turn stimulates economic activity and boosts exchequer receipts. However, when house prices are falling, homeowners can reduce their spending as they begin to fear that their debt burden from their mortgage will outsize the value of their property, thus restricting economic activity (Bank of England, 2018a; Zhu, 2005). Given that these factors are primary drivers in policy setting, it is logical then that governments, central banks and other regulatory bodies are interested in monitoring the state of the market over time.

The key role of government is to foster a sustainable, responsible and beneficial economy for citizens and businesses in the country they govern. With shelter being one of Maslow's fundamental human physiological needs, it is critical that governments ensure their citizens have access to safe, affordable housing, in order to stimulate a healthy and stable economy (Maslow, 1943). In more recent times, it has become increasingly difficult for young, first-time buyers to get a foot on the property ladder (McKee, 2012). This is primarily due to a shortage of available housing driving prices to all-time highs, as demand increases with population (Conefrey, Staunton, et al., 2019).

Governments around the world have been attempting to tackle this issue by

stimulating homebuilding through grants, competitive loans and tax breaks for developers, and by offering tax-efficient vehicles and duty relief for first-time homebuyers. It is critical that these policymakers keep a watchful eye on the state of the property market, in order to judge the extent, balance and scale of the supports they are putting in place. Increasing support on the buy side must be balanced with property completions, otherwise the situation will only worsen through a widening of the gap between supply and demand; increasing prices further.

Central banks also have a vital role in steering the economy, by setting monetary policy. Their interest in property prices cannot be understated, given the considerable economic impact that property has on the economy, as discussed previously. Home prices rising too quickly feeds into inflation, which most central banks are mandated to control, while a housing crash generally leads to recession and subsequently, an increased unemployment rate (Bank of England, 2018a).

The primary tool used by central banks to ease or tighten monetary policy is the overnight policy interest rate. This is the rate that a depository pays to borrow money overnight from another depository, in the domestic currency. While the immediate pass-through impact of an increase in the policy rate on mortgage rates varies by country, according to the mix of fixed-rate and floating-rate mortgages, it remains the primary determinant of residential mortgage rates (Hess and Holzhausen, 2008). This strong relationship between residential mortgages and monetary policy tools leads to central banks needing to be cautious in ensuring that their actions do not result in a crash or a bubble and thus, the up-to-date monitoring of market trends following a change in the policy rate is of keen interest to the policy makers.

2.4.4.1 Modelling viability

Naturally, the primary set of statistics used by both governments and central banks will be those produced by the national statistics board of the country in question. For property prices, this will typically be either a hedonic regression model, or a repeat-sales model, as discussed previously. Thus, the same key issues will apply; namely that the house price index is produced with a considerable lag, which makes it difficult to see the impact of monetary and fiscal policy in real-time.

Central banks often turn to alternative, more timely measures of critical economic indicators as a supplementary tool, while still viewing the indices published by the statistics bureau as the *gold standard*. The Federal Reserve of the United States of America, as an example, monitor a variety of different property price models to obtain a diversified, aggregate view on the market (Rappaport et al., 2007). Each of the models offer advantages and disadvantages, however, it demonstrates that there is a use-case for alternative house price indices, even among the highest level of market stakeholders.

Owing to their position of power and significance, these parties have the resources to obtain the necessary data and implement any type of statistical model they desire, however, like other market observants, they may have to accept a reporting delay, depending on what kind of information they require. All of the property price models discussed are viable for both government and central banks and, as shown, they typically tend to monitor a varied assortment of data sources, increasingly including in-house statistical modelling and machine learning (Bholat, 2015).

2.5 Chapter summary

Property price models vary substantially in both their methodology and their composition, without any general consensus on which model produces the most accurate or precise house price index. The commonly used hedonic regression and repeat-sales models each have distinct benefits and drawbacks, and their viability depends on data, resources and required timeliness. Mix-adjusted median models offer a less resource-intensive method of generating a price index, however, they have been given less attention and forethought than the methodologies mentioned prior.

Due to the number of stakeholders in the housing market, each of which would benefit from a more punctual view on housing trends, it is important to explore viable alternatives to the status quo; notably those which can improve upon drawbacks in the existing models. It is not necessarily the case that these novel methodologies must serve as a total replacement for the established tools, particularly given

the lack of any convention; they could be adopted as complementary methods of evaluating the state of the market.

Chapter 3

Initial Work: A robust house price index using sparse and frugal data[1]

As discussed in the previous chapters, property price index models are of keen interest to a number of market stakeholders. Despite this, the most commonly used house price index methodologies today have a number of unaddressed shortcomings, which result in potential measurement inaccuracies, the need for an extremely rich and expansive dataset and an untimely publication of the resulting index.

In this chapter, a new mix-adjusted property price index model will be introduced, one which is designed to leverage the underutilised presence of spatial autocorrelation in housing. The goal of this is to replace the need for a rich set of attribute data on each property transacted over the period of interest, resulting in a more transparent, reproducible and automated methodology.

By matching transacted properties to neighbours in prior time periods, it is posited that the high likelihood of those neighbours sharing similar price-influencing characteristics will result in them being broadly comparable across different time periods, thus allowing a measure of price trend to be calculated. This theory has strong basis in existing literature, as discussed in Chapter 1.

The *GeoPrice* index will operate on a publicly available dataset which is both sparse and frugal; the total number of transactions are low and there is only the most minimal set of information on each property transacted. In order to benchmark the performance of the model, comparisons will be made against the results of the

[1] This chapter is adapted from *A robust house price index using sparse and frugal data* (Maguire, Miller, et al., 2016), with additional contextual information added.

official Irish Residential Property Price Index, produced by the Irish Central Statistics Office (CSO).

3.1 The Irish Residential Property Price Index (RPPI)

The Irish Residential Property Price Index (RPPI) is produced by the Central Statistics Office (CSO) on monthly data samples using a hedonic regression model. The data driving this model is sparse; with a typical monthly transaction volume of approximately 2,200 in our sample period[2]. As a result of this, price indices are not available on a county-by-county or regional level; the only geographic sub-indices are *Dublin* and *Rest of Ireland*.

The primary data source leading into the production of the RPPI is mortgage returns data. All lending agencies in Ireland must file a mortgage return for any property sold in the country which is funded by a mortgage, whether wholly or in-part. These returns must be reported by lenders on a monthly basis to the Department of Housing, Local Government and Heritage, which makes this protected dataset available to the CSO for the purposes of producing their house price index (O'Hanlon, 2011). As a result, the reporting and administrative lag associated with this process directly feeds through to the production of the RPPI.

The advantages of using mortgage return data in the production of the RPPI is the ability to compare properties with a high level of similarity to one-another, through the use of the rich data on each property included in the mortgage return; number of bedrooms, number of bathrooms, floor area, total plot area, age of property etc. However, the use of mortgage data comes with a number of key disadvantages. The number of errors in mortgage returns is exceptionally high; estimated at approximately 68% by the CSO, resulting in a need for imputation (O'Hanlon, 2011). Additionally, the reliance on mortgage returns for the provision of property attributes means that cash sales must be excluded from the model entirely.

During times where the share of sales completed using a mortgage is very high, this may have a negligible impact on the sample. However, in periods where lending is restricted due to economic uncertainty and/or low levels of mortgage affordability

[2] Data was sampled from 2010 to 2015, inclusive.

in the population, the absolute number of mortgage approvals tend to fall, while the number of cash transactions remain similar; leading to an increased share of cash transactions in the monthly sale sample. For example, in the years succeeding the *Global Financial Crisis* of 2008, the share of mortgage transactions plummeted from 88% to 50%, owing to the introduction of stricter regulatory controls on mortgages (Dalton and Moore, 2014).

The primary issue with the varying share of cash sales not being accounted for in the RPPI is the bias it introduces on the sample. If the distribution of cash sales and mortgage sales were near identical, this issue could be disregarded, however, cash transactions typically involve cheaper properties, whereas more expensive houses are usually purchased via a mortgage, by necessity. This sample bias is not taken into account by the model and thus the results may be somewhat skewed, as a result. Furthermore, the variability in the share of cash transactions across different time periods means that the bias is changing on a monthly basis and thus, the effect it has on the index is not even consistent over time (Dalton and Moore, 2014).

3.1.1 Methodology

The Irish Residential Property Price Index uses a 12-month rolling time-dummy hedonic regression model, the methodology of which is discussed in Section 2.1. However, due to the low volume of monthly transactions mentioned previously, the published index takes a three month rolling average of the raw index values output by the model (O'Hanlon, 2011). While this artificially increases the smoothness of the index, it increases the amount of time needed for market changes to propagate to the RPPI, which is already a significantly lagged publication. There is also a possibility that this smoothing process could convey a false sense of conviction in the index, given that the un-smoothed index is not made readily available to the public, and this procedure is not plainly communicated to users of the data.

As mentioned previously, imputation is also used by the CSO to interpolate any attributes from the mortgage return which are missing or deemed to be clearly and obviously erroneous or implausible. While this is a reasonable approach which is

utilised in many hedonic regression house price models[3], there are a number of issues in applying it to Irish mortgage returns data. Given the low transaction volume mentioned previously, in addition to the fact that the majority of mortgage returns have at least one error, the pool of *valid* data points from which imputation can be derived is likely to be *~1,000 to 1,500* per month. Even if the imputation pool is extended to a rolling one or two year dataset, the total volume is still relatively small and the risk of stale data permeating through the index is heightened.

According to (O'Hanlon, 2011), the lack of the Residential Property Price Index to account for *"quality of neighbourhood analysis"* and inability to leverage the *"explanatory power of location coefficients"* through geospatial stratification of the hedonic regression model is *"undoubtedly the most serious weakness"* of the index. The failure to factor this characteristic of properties is due to insufficient and unreliable address details provided through mortgage returns data, which makes proper segmentation of localities infeasible.

3.2 *GeoPrice*: A sparse and frugal property price model

3.2.1 The Property Price Register dataset

As an alternative to mortgage returns data, stamp duty returns are maintained by the *Property Services Regulatory Authority* (PSRA) and are publicly available online via the Property Price Register (Property Services Regulatory Authority (IE), 2024). This service contains details of every property sale transaction in the Republic of Ireland from January 1st, 2010 onward, however, the only attributes available for every sale record are: date of sale, sale price and address. This dataset has the advantage of a lower lag from reporting to publication time, with a typical lead time of ten days, yet there are also a number of drawbacks. Aside from the frugality of the property attributes, there are cases where the delay in reporting of a property sale can stretch several months beyond the actual sale date. Furthermore, transactions where an explicit sale does not occur, for example, an inheritance, are included within the dataset without any reliable filtering method. These transactions may cause a small

[3] e.g. in the Office for National Statistics UK House Price Index (ONS, 2023b)

distortion in any analysis produced upon them, owing to the fact that these transactions typically have a reported value significantly lower than the market value of the property in question.

These drawbacks are not sufficiently problematic to rule out the potential use of the dataset however; it seems reasonable to assume that the transactions with delayed reporting are randomly distributed, such that this effect does not introduce a bias on the monthly sample. The type of property being sold should not affect the probability of the that sale being a delayed transaction, as delays are typically due to administrative issues or human error. Furthermore, the scope of these delays should be bounded by legal disincentives for late filing of stamp duty returns. By law, filings which are late by more than forty-four days are subject to penalties, with these increasing over time [4]. As such, the number of buyers who do not comply with these regulations should be small and thus the impact of delayed reporting on the index should be limited.

A similar argument can be made with regard to the non-sale transactions; as inheritance is typically associated with the passing of the property owner, there is no reason to believe that this effect will be more pronounced in one time period than another, or in one location than another. Given that these effects can be reasonably expected to be random, they should cancel out across a sufficiently diverse dataset, owing to the law of large numbers.

As an initial, basic analysis, we can look at the distribution of the monthly property sale pools in the Property Price Register data (see Figure 3.1). The variability in transaction volume from month to month is significant, as shown in Figure 3.2. This inconsistency in the number of transactions likely means that months where volumes are particularly low will be substantially more noisy than months where the number of sales is ample. Prior to any filtration being performed, the month with the lowest volume was January 2011, with 1,037 transactions, while the month with the highest number of sales recorded was December 2014, with 7,523 transactions; over 50% more than the next closest time period.

It is clearly evident from Figure 3.2 that transaction volumes are highly seasonal;

[4] See: https://www.revenue.ie/en/property/stamp-duty/paying-the-duty/late-filing-and-paying.aspx.

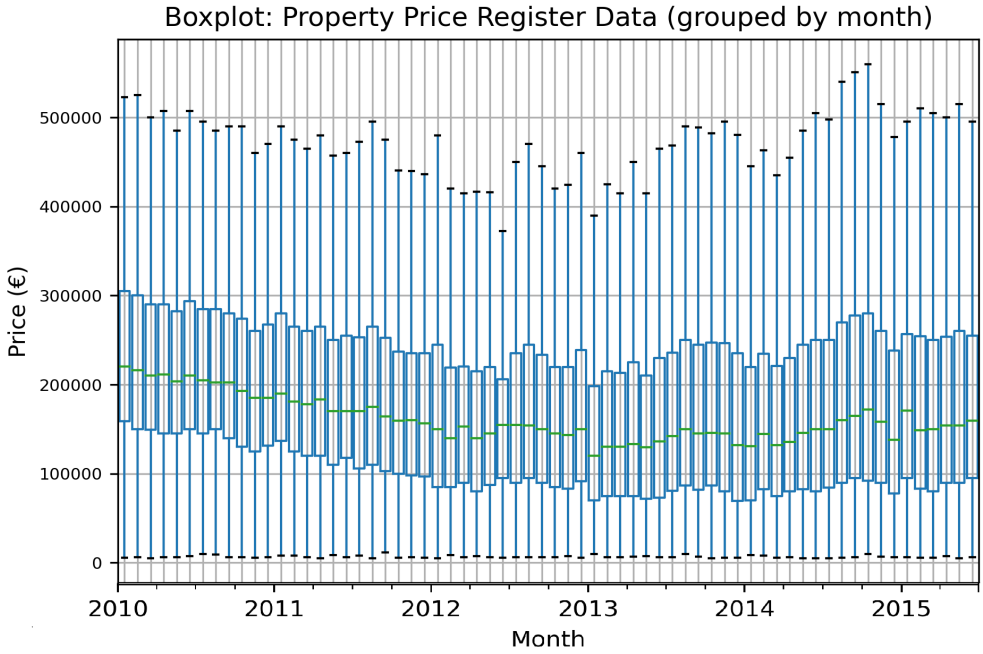


FIGURE 3.1: Property Price Register Price Distribution from 01-2010 to 06-2015 (inclusive), grouped by month.

the number of sales tends to dramatically increase towards the end of each year, then drops off sharply. Indeed, the seasonal analysis in Figure 3.3 shows that strong multiplicative seasonality with annual periodicity can be detected within the monthly transaction buckets.

While neither the mean nor the median are particularly insightful tools in analysing property prices, owing to the amount of volatility produced by unhandled shifts in the housing stock of each monthly basket, they nevertheless can serve as a naive benchmark and reference point for smoothness when comparing property price index models later in this chapter. A primitive house price index which takes the mean and median price of each monthly bucket is shown in Figure 3.4.

Figure 3.5 also shows a potential correlation between the simple mean monthly price and the transaction volume by month. As discussed, transaction volumes spike towards the end of each year and rapidly drop off once the new year begins. This idiosyncrasy of the Irish property market appears to permeate through to the mean price, with a seasonal drop observed in the first three months of the year. Prices also appear to spike in the middle of the summer, with a similar uptick in volume measured in the transaction seasonality for the same period. Interestingly however,

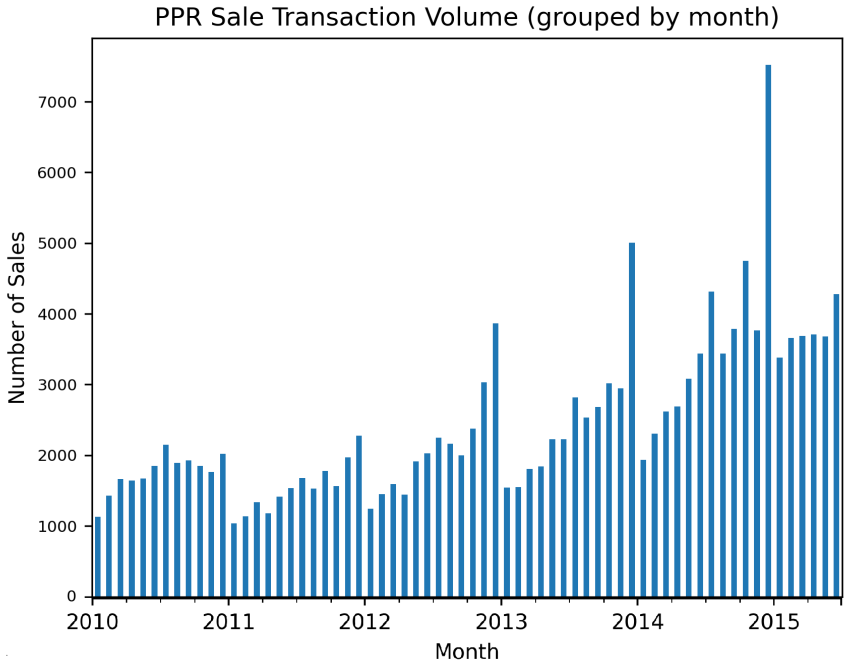


FIGURE 3.2: Property Price Register Data Volume from 01-2010 to 06-2015 (inclusive), grouped by month.

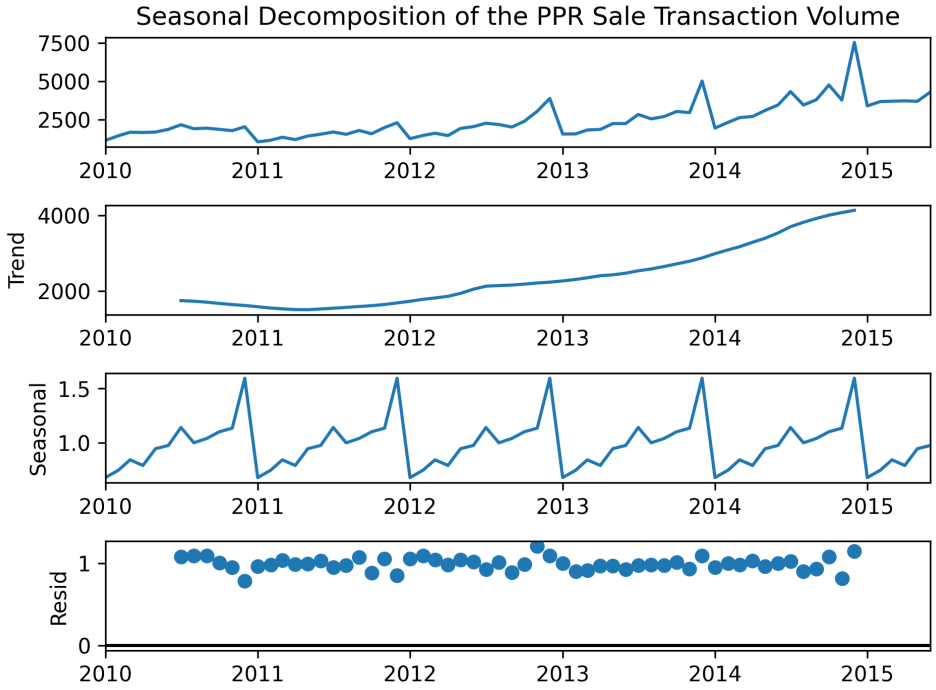


FIGURE 3.3: Property Price Register Data Volume Seasonality from 01-2010 to 06-2015 (inclusive), grouped by month.

the large upturn in sale volumes observed towards the end of each year does not materialise as a corresponding surge in the mean price.

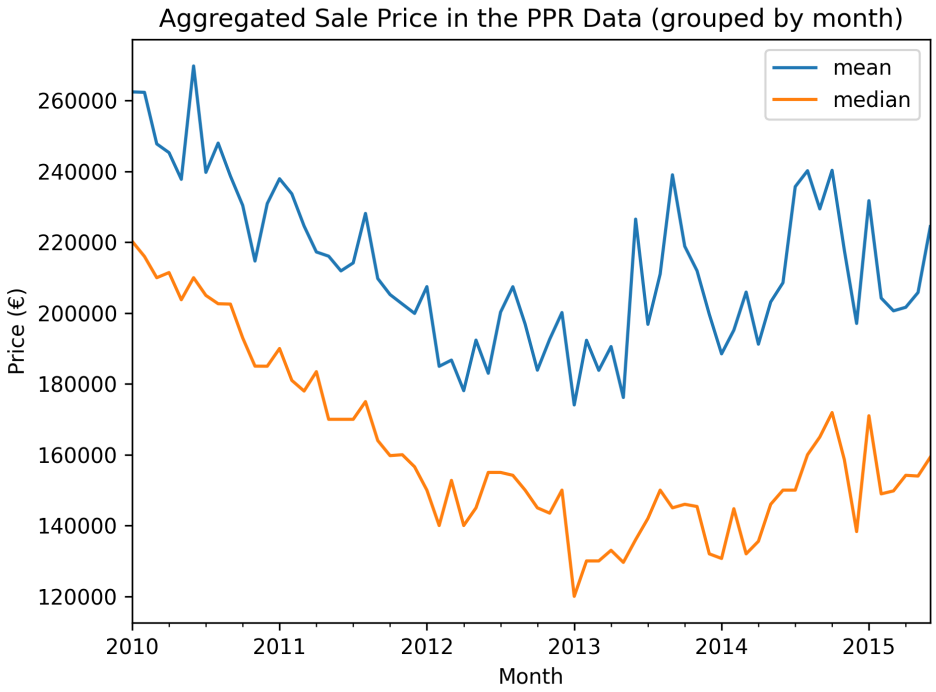


FIGURE 3.4: Property Price Register Mean/Median Price from 01-2010 to 06-2015 (inclusive), grouped by month.

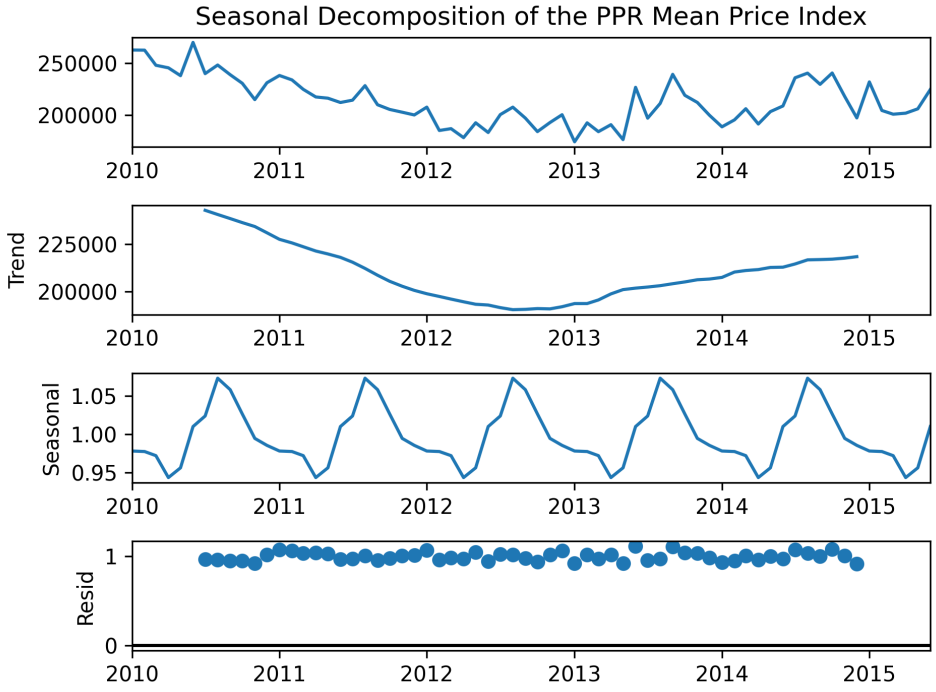


FIGURE 3.5: Property Price Register Mean Price Seasonality from 01-2010 to 06-2015 (inclusive), grouped by month.

3.2.1.1 Limitations

The primary challenge with using the Property Price Register data is inconsistent address formatting. Currently, the Property Price Register makes the address available via a single text field, with the address directly taken from the stamp duty return. No constraints are enforced on the clarity and precision of the submitted address and no validation, cleaning or formatting is done on the raw text; it is uploaded as-is. Unfortunately, many addresses in Ireland are poorly described and difficult to identify. In recent times, Ireland has adopted a unique postal code system, *Eircode*, and while this does exist on some of the newer property sale transactions in the Property Price Register data, it does not exist historically and is not guaranteed to be present. This makes this data field impossible to use when attempting to test a model on a historical sample period.

In order to perform our analysis and fit our index model, we collected data from January 2010 to June 2015 (inclusive). To adequately handle the address issue, we used Google Maps mapping system, widely considered to be best in class, to match the addresses in the Property Price Register dataset to precise GPS co-ordinates. Due to the rate limits on the Google Maps geo-coding API, we were only capable of querying *2,500* records per day. Thus, collection of co-ordinates for our sample set had to be carried out over a total period of approximately two months. While it was not possible to accurately match every address to a pair of GPS co-ordinates, we were able to achieve a hit rate of approximately 90% of reported transactions, for a total of 147,635 unique property sales. The data was split into monthly sets for our model, to allow a direct comparison with the monthly RPPI publication.

The Central Statistics Office have claimed that the lack of property attributes in the stamp duty data from the Property Price Register makes the dataset impracticable for use in a property price model. According to O'Hanlon, 2011, this is due to the vague data not offering any viable opportunity to perform mix-adjustment on the sample set. However, our analysis will investigate the feasibility of producing a house price index of equal-or-greater robustness to the RPPI using the public Property Price Register data, without any supplementary property attribute information.

3.2.2 Methodology

The *GeoPrice* index is designed to leverage the spatial auto-correlation effect of properties at each stage of the methodology. By capitalising on the inherent similarities shared by neighbouring properties, the index intends to implicitly match similar properties across multiple time periods by looking at their proximity to one-another. As such, it is no longer essential to explicitly model and adjust for the attributes of each property; the price trend of houses deemed likely to be homogenous through being neighbours can serve as comparable samples across different periods of time, thus allowing price trend to be estimated.

3.2.2.1 Stage One: Filtering

As an initial filtering stage, properties where additional transactions have been recorded within 100 meters in a forty-eight hour time period (i.e. $\pm$ 24 hours) of the sale are discarded from the dataset. The reasoning for this is two-fold; firstly, analysis of the stamp duty returns data demonstrated that entire housing estates or apartment blocks are frequently sold in bulk within the same time period, which significantly distorts the average price of transactions in that given period of analysis. Secondly, duplication errors were occasionally observed on the Property Price Register portal, where the same property was entered multiple times. This filtration method results in the stripping out of both of these data contaminants, ensuring less noise is recorded in the final model output.

Algorithm 1 formalises the logic underpinning the stage one filtering process. Applying this data purification process resulted in the loss of approximately 14.4% of the dataset, reducing it to a total 126,444 transactions over our sample period.

3.2.2.2 Stage Two: Proximity Voting

As demonstrated in Section 3.2.1, property price index models which are based around the median are subject to dramatic volatility due to fluctuations in the composition of the the monthly basket. As an example, if a locality such as Dublin was to experience twice the typical number of sales in a given month, being a region in

[5]See: (Robusto, 1957) for the haversine distance formula.

Algorithm 1 Sparse and Frugal Model: Stage One - Filtering

property_set: a set of all properties in the property data sample
records_to_remove: a placeholder set, for properties marked for deletion

$THRESH_{DIST}$: 100
$THRESH_{TIME}$: ONE_DAY

procedure EXCLUDE_NEIGHBOURS(*properties*, *record*)
  for all p_i in *properties* **do**
    if TIME_DELTA(p_i[sale_date], *record*[sale_date]) $\leq THRESH_{TIME}$ **then**
      if HAVERSINE_DIST(p_i, *record*) $\leq THRESH_{DIST}$ **then**
        $records\_to\_remove \leftarrow records\_to\_remove \cup \{p_i\}$
      end if
    end if
  end for
end procedure

for all p_i in *property_set* **do**
  EXCLUDE_NEIGHBOURS(*property_set*, p_i)
end for

$property\_set \leftarrow property\_set \setminus records\_to\_remove$

▷ HAVERSINE_DIST is a procedure which computes the haversine distance [5] between two properties' GPS co-ordinates, in meters.
▷ TIME_DELTA is a procedure which computes the absolute time difference between two timestamps.

which the median price of the region is above the national median, this would drag the national median upwards, despite not representing a genuine increase in price.

For example, Prasad and Richards, 2008 found that variations in the monthly observed housing stock between higher and lower valued parts of cities led to substantial volatility in the unadjusted median price of their dataset of 3.5 million property sales in Australian cities. Furthermore, US realtors have reported that a seasonal effect pushes American property prices higher in the summer, due to the purchasing habits of families with children tending to buy during the school year holidays. These homeowners typically purchase more expensive properties than the median, thus exerting an upward pressure on median-based price indices in months affected by this seasonal pattern (Prasad and Richards, 2008). Indeed, a similar summer seasonal effect was observed in the Irish Property Price Register dataset, as seen in Figure 3.5.

Thus, in order to design a robust house price index, it is necessary to develop methods which can control for the mix of properties feeding into each monthly sample set. In other words, we wish to extract a subset of each monthly sample which is more representative of the mix of housing stock in the market being studied, as a whole. In this analysis, our focus is on the use of geospatial stratification to adjust the mix, thus, we aim to select a sample which is maximally spatially autocorrelated with the historical mix, ensuring that the distribution of transactions across different regions of the country remain stable throughout each month of the data. This would serve to minimise the bias contaminating the index and increase the comparability of the monthly observations.

Spatial autocorrelation tends to be present in properties for a number of reasons. Firstly, properties which lie near to each-other are often members of the same unit of homes, or a component of a large development of contiguous housing (Mar Iman, 2001). As a result of this, homes in close proximity tend to have a number of characteristics in common, for example, similar floor area, age of dwelling, design features and property type (Ismail, 2006). The primary driver of this similarity is that blocks of housing tend to be developed at the same time. These housing blocks are also more representative of the typical house than, for example, a large, secluded detached property which was developed as a single, custom unit (Gillen, Thibodeau,

and Wachter, 2001). Therefore, these properties with high levels of spatial autocorrelation tend to compose a large share of the sample mix and thus exert more influence on the house price index.

Additionally, an often understated advantage of leveraging geospatial proximity in property modelling is an implicit handling of environmental factors. As discussed in Section 2.1.1, environmental characteristics including quality of schools, public transport links, green space and crime rates, among others, have notable influence and explanatory power over a property's value. Where the majority of housing models tend to omit these factors due to the difficulty of acquiring and integrating data pertaining to them; geospatial proximity can be exploited to incorporate these factors, owing to the fact that proximate properties will share the same environmental factors, to a high degree (Ismail, 2006).

In light of these motivations, we designed a system to increase the level of spatial autocorrelation in our sample, by filtering out 10% of the dataset which were deemed to be the least representative properties. In order to select this portion of transactions, a single transferable voting system was introduced, where properties in the historic sample set vote for their nearest neighbour in the set of transactions for a given month. Algorithm 2 formalises this process.

As the votes for each month are cast based on the distribution of the entire historic sample set, this should ensure that the number of properties from each region remains broadly stable over time. For instance, if Dublin was to experience twice as many sales as usual in a given month, a number of these sales would be stripped from the dataset, as the voting system is based on the proportions set by the historic data. As each property in the historic data can only cast one vote, it is not possible for all of the additional sales in Dublin to be elected. Furthermore, this method should increase the likelihood that homes with similar property type and environmental factors remain generally proportional across months, owing to their correlation among proximate properties.

As a worked example of this algorithm, assume month X is being used as the electorate for candidates in month $X + 1$ (in practice, multiple months will be used as a electorate, but we simplify here for the sake of example). Again, for the sake of simple example, assume that month X has a total of $10,000$ property sales, while

Algorithm 2 Sparse and Frugal Model: Stage Two - Proximity Voting

property_set: a set of all properties remaining following Stage One
elected: a placeholder mapping of months to their set of elected properties
votes: a placeholder mapping of properties to their total number of votes
trim_ratio: 0.1

M_i: the month i of the sample period
P_i: the properties in *property_set* for M_i

procedure VOTE(*candidates*, *voters*)
  for all v_n in *voters* **do**
    $c_n \leftarrow NN(v_n, candidates)$
    $votes[c_n] \leftarrow votes[c_n] + 1$
  end for
end procedure

procedure ELECT(*candidates*, M_i, *voter_cardinality*)
  $election\_thresh \leftarrow \frac{voter\_cardinality}{(1-trim\_ratio)|candidates|}$
  for all c_n in *candidates* **do**
    if $votes[c_n] \geq election\_thresh$ **then**
      $excess \leftarrow votes[c_n] - election\_thresh$
      $elected[M_i] \leftarrow elected[M_i] \cup \{c_n\}$ ▷ Add c_n to the set of elected for M_i
      $candidates \leftarrow candidates \setminus \{c_n\}$ ▷ Remove c_n from *candidates*

      $\hat{c_n} \leftarrow NN(c_n, candidates)$ ▷ Find the NN to c_n in *candidates*
      $votes[\hat{c_n}] \leftarrow votes[\hat{c_n}] + excess$ ▷ Add the *excess* to $\hat{c_n}$
    end if
  end for
end procedure

procedure ELIMINATE(*candidates*)
  $e \leftarrow MIN(candidates)$ ▷ Get candidate with least votes
  $candidates \leftarrow candidates \setminus \{e\}$ ▷ Remove e from candidates

  $\hat{e} \leftarrow NN(e, candidates)$ ▷ Find the NN to e in *candidates*
  $votes[\hat{e}] \leftarrow votes[\hat{e}] + votes[e]$ ▷ Redistribute e's votes to $\hat{e}$
end procedure

for all M_i, P_i in *property_set* **do** ▷ Iterate over each month of transactions
  $historical\_transactions \leftarrow \sum_{n=0}^{i-1} P_n$ ▷ $\sum$ here refers to set concatenation
  $historical\_cardinality \leftarrow |historical\_transactions|$

  VOTE(P_i, *historical_transactions*)

  while $P_i \neq \emptyset$ **do** ▷ Repeat until all of P_i have been elected/eliminated
    ELECT(P_i, M_i, *historical_cardinality*)
    ELIMINATE(P_i)
  end while
end for
▷ MIN is a procedure which returns the property p from the *candidates* argument for which $votes[p]$ is the minimum.
▷ NN is a procedure which returns the nearest neighbour to the first argument, among the set of data points passed as the second argument.

month $X+1$ has a total of $1,000$ transactions. Given a 10% trim ratio, this will result in each property needing a total of at least $11.\dot{1}$ votes to be elected to remain within the sample for month $X+1$. Firstly, each of the $10,000$ properties will cast their vote for their nearest neighbour in the $1,000$ transactions within month $X+1$. Any properties which exceed the threshold of votes will be elected, and their excess votes, if any, will be distributed to their nearest neighbour. Then, the property with the least votes will be eliminated and their votes will be distributed to their nearest neighbour. This process continues until all properties have been either eliminated or elected.

3.2.2.3 Stage Three: Localised stratification

Although mix-adjustment aids substantially in increasing stability and lowering volatility in the median house price measure, there is scope for further improvement through additional analysis of the transaction set. By solely considering the median of each monthly sample, information on the distribution of prices above and below the median value is effectively ignored. If the shape of this distribution fluctuates between months, this detail should be exploited in order to enhance index stability further.

In the case where the various regions present in the property data have distinct medians, with diverging price action, this issue becomes even more pronounced. As an example, the national mix-adjusted median house price at the start of 2015 was €180,000. Owing to the fact that capital cities tend to be more expensive to buy property in, a substantial proportion of Dublin homes sold for a value above this national median (circa 85% of Dublin transactions). If we were to solely look at the median, any price action particular to Dublin would have scant impact on the national median price, due to being mostly above it already.

On the contrary, localities in the data which share a median value which is very similar to the national median price have a disproportionate effect on the national index. Owing to their proximity to the median value, they contribute too much information to this measure, while areas which are on each end of the price spectrum contribute too little information, despite potentially having a generous share of the property mix, particularly in the case of capital cities such as Dublin.

As such, it is critical to disaggregate the transaction set geospatially in order to construct a representative house price index, where areas with differing medians are contributing information to the national index proportionally (Goh, Costello, and Schwann, 2012). Evidence from the Australian housing market reveals a marked difference in house price behaviour between different metropolitan areas; information which would be mostly lost through the application of the simple median (Costello, Fraser, and Groenewold, 2011; Hatzvi and Otto, 2008).

Stratification is one potential method of achieving this. Prasad and Richards, 2008 proposed a novel algorithm for stratifying an Australian dataset. Suburbs were grouped together based on the long-term average price level of dwellings in said regions, with a weighted average of the medians of each stratum composed the national index. They concluded that this measure of property prices was a significant improvement on the unstratified median measures, and achieved a high level of correlation with hedonic regression models in the same region.

A similar case study was carried out by McDonald, Smith, et al., 2009 on New Zealand property data; applying an analogous approach to that of Prasad and Richards, 2008. The results achieved by this investigation concurred with their conclusion; their novel stratified measure of property prices was highly correlated with the regression-based *QV Quarterly House Price Index*, one of the primary house price indices observed in New Zealand [6]. However, they further concluded that this measure of property prices was capable of reporting in a much more timely fashion; a key problem faced by the commonly used housing models, which we discussed in depth in Section 1.2.

A limitation of the model used in both of these studies is that the data has been segmented into arbitrary strata; the boundary between a property being in one stratum or another is entirely arbitrary. There is no guarantee that the strata chosen for these studies are optimal for volatility reduction, nor is it certain that they are the most representative delineations of the market in their respective regions. It is possible that these strata may drift over time, due to environmental factors in a region changing, such as improvement in schools, or new green space developments

[6] See: https://www.qv.co.nz/price-index/

increasing the value of properties in said locality (Luttik, 2000). This concern is not factored into the technique proposed by Prasad and Richards, 2008.

In order to solve this issue, we propose removing the arbitrary boundaries between strata by giving each and every property its own local base. In order to illustrate, first assume that two months of properties have been selected, a stratification base month, and a month where the price change is to be evaluated, called the *current* month. The sale price of each property in the current month is divided by the price of the nearest property in the base month; thereby giving a set of price ratios. We then take the median of those price ratios, giving our stratified, filtered, mix-adjusted median price index measure. Algorithm 3 formalises this process.

Algorithm 3 Sparse and Frugal Model: Stage Three - Localised Stratification

elected: a mapping of months to their set of selected properties from Stage Two
month_ratios: a placeholder mapping of months to a set of price ratios
month_changes: a placeholder mapping of months to their computed price change, relative to the stratification base

M_B: the chosen stratification base month
P_B: the properties in *elected* for M_B

M_i: the month i of the sample period
P_i: the properties in *elected* for M_i

procedure STRATIFIED_RATIOS(*property_sales*, M_i)
  for all p_n in *property_sales* **do**
    $\hat{p_n^B} \leftarrow NN(p_n, M_B)$ ▷ $\hat{p_n^B}$ is p_n's nearest neighbour in M_B

    $month\_ratios[M_i] \leftarrow month\_ratios[M_i] \cup \{\frac{sale\_price(p_n)}{sale\_price(\hat{p_n^B})}\}$
  end for
end procedure

for all M_i, P_i in *elected* **do**
  $STRATIFIED\_RATIOS(P_i, M_i)$
  $month\_changes[M_i] \leftarrow MEDIAN(month\_ratios[M_i])$
end for

▷ NN is a procedure which returns the nearest neighbour to the first argument, among the set of data points passed as the second argument.
▷ *sale_price* is a procedure which returns the sale price for the property passed as an argument.

As a worked example, suppose a house is sold in Wicklow for €240,000, in January 2015. Suppose we select the stratification base month to be November 2014.

Then we would find the nearest neighbour of the house sold in Wicklow, in November 2014's sale records. Let's assume this to be a house worth €200,000. Then, the price ratio of these properties would be $\frac{240{,}000}{200{,}000} = 1.2$, which would imply a change of $+20\%$, from November 2014 to January 2015. We would take the median value of all of these price ratios, in order to calculate the overall change from the base month to the selected month.

Through this method, all areas contribute information to the price index, as we have stripped out the oversized impact of properties close to the national median on the index. By calculating our model on localised, stratified price ratios, market data coming from any property in the transaction data has an equal opportunity to impact the index as any other, reducing the model's volatility.

3.2.2.4 Stage Four: Leveraging multiple base months to reduce volatility

To push our model even further, we can enhance the stability of the index by leveraging multiple stratification base months, rather than just a single one. As an example, the house price index for January 2015 could be derived by computing the price ratios using December 2014, Novemeber 2014, October 2014 and so forth, each as the stratification base month. Table 3.1 displays the results of running Stage Three of the model on January 2015, using the prior six months as different stratification base months.

In order to generate our final index, we calculate the monthly change using every available historical stratification base and take the average monthly change implied by each of those values. We will explore the results of each of these stages of the algorithm in Section 3.4, including a comparison with the CSO's RPPI index, however, we first must explore methods of comparing the performance of distinct house price index models.

3.3 Measuring robustness through smoothness

The measurement of robustness is of key importance to research in the area of property prices, owing to the fact that no *gold standard* of measurement exists. Despite

TABLE 3.1: Stage Three Model: Index change from December 2014 to January 2015

Base Month	Jul 2014	Aug 2014	Sept 2014	Oct 2014	Nov 2014	Dec 2014
Jan 2015 [a]	+5.7%	+3.9%	+3.8%	+2.7%	+2.2%	+6.0%

[a] Values represent the month-over-month change from Dec 2014 to Jan 2015, using different stratification base months.

this, methodology surrounding evaluation of house price index robustness has received little attention in the literature (Goh, Costello, and Schwann, 2012).

How does one distinguish a *good* house price index, from a *bad* house price index? Chandler and Disney, 2014 found that it was surprisingly difficult to identify what a property price index is even intended to measure. The UK's Office for National Statistics states *'the aim of the ONS House Price Index is to measure the change in the average house price for owner-occupied properties in the UK'*. On the other hand, *Nationwide*, a large British mortgage lender, produce their own mix-adjusted median house price index. They state that their index measures *"the price for a 'typical' house"* (Nationwide, 2024).

These house price models are therefore attempting to measure different things, with the ONS seeking to measure the change in the average house price, while the Nationwide index looks to measure the price movement of a typical house. This subtle disparity is relevant to our prior discussion in Section 3.2.2.3, where an index which only considers the median value is simply measuring the movements around a *typical* property; largely ignoring the behaviour of the market behaviour above and below what is considered *typical*.

However, in cases such as the ONS house price index, what does the *average* house price actually mean? The subtle, yet significant differences between house price index methodologies, definitions and objectives makes it difficult to assess the accuracy and robustness across the various commonly used models. The *gold standard* of house price trend is unobservable, as it would require measurement of the entire housing stock in every time period of interest (Goh, Costello, and Schwann, 2012). While the model produced by Wallace and Meese, 1997 assumes that the true

house price trend can be proxied by observing the median house price, a large body of literature argues against the use of the median alone (Case and Shiller, 1987; Goh, Costello, and Schwann, 2012; Hansen, 2009).

Goodness-of-fit statistics are one of the tools commonly employed for assessing the accuracy of models, used by Case and Szymanoski, 1995 and Prasad and Richards, 2008 to assess their proposed house price indices, for example. On the other hand, Sommervoll, 2006 argues that this metric can be highly misleading, due to the risk of overfitting. This is particularly concerning in the case where indices are disaggregated into a very granular level of detail, or when models are fit to sparse datasets. It is possible for substantial inaccuracies in the measurement of market behaviour to occur despite a high R^2 or $t - value$ reading, as powerful models can be easily tuned to produce highly convincing results when strata have a small number of samples.

The main drawback in goodness-of-fit measures is the lack of penalisation for model complexity. In other words, these metrics should take into account the number of free parameters which the model has available to tune. A given model is theoretically capable of achieving an arbitrary goodness-of-fit measure on the sample set simply by dramatically increasing the number of model parameters available to it. This is a particular issue in recent times, owing to the increasing prominence of complex machine learning models which can have thousands, or even hundreds-of-thousands of adjustable parameters (Domingos, 2012).

A potential alternative to goodness-of-fit, suggested by Goh, Costello, and Schwann, 2012, is cross-validation. They propose adopting a system whereby 75% of the transactions are selected from the sample at random, and the other 25% of the sample is used to verify how well the model has been fit. The closer the match between the training set and the test set, the more robust the model. There are some issues with this metric, however. Firstly, it cannot be said that a single instance of cross-validation guarantees a good fit; it is possible that two random halves of a dataset produce similar results by chance, where other partitions of the sample may have given contradictory results. In order to ensure the accuracy of this metric, the cross-validation would need to repeated a number of times, with the *mean agreement* of the partitions in each test cycle taken as the benchmark for a good fit.

Furthermore, the issue with this method of assessing model robustness is that trivial models which ignore the input and produce near-identical values on every run will generate a staggeringly strong result using this metric, as results will be highly consistent, yet they hold no information whatsoever on what is being attempted to be measured. In practice, the most robust house price index, is the smoothest one. The driving nature behind this claim, is that the property market is historically known to have a cyclical pattern; during times of market strength, house values surge consistently upwards for months or years in succession (Agnello and Schuknecht, 2011; Leung and Tsang, 2013). During recessionary periods, prices begin to move the opposite direction, and will consistently drop month-to-month for an extended period of time (Chen, Kawaguchi, and Patel, 2004). Indeed, Englund and Ioannides, 1997 found statistical evidence of strong autocorrelation in the housing market of no less than fifteen *OECD* countries. This strong autocorrelation suggests that price action in the current month should have predictive power on price action in upcoming months, which has also been observed by Case and Shiller, 1988, among others. In other words, the property market trend is expected to be smooth and carry momentum forward, rather than oscillating up and down between months.

In contrast, house price models experience various degrees of sampling errors, which jump around and frequently flip direction on a month-to-month basis. These sampling errors, unlike the housing market, don't have a particular momentum or trend. This would explain the reasoning behind the CSO utilising a three-month smoothing technique to reduce the impact of these errors, as they tend to partially cancel out in successive months. As a result of this divergence in frequency, smoothness acts as a strong metric for robustness and reliability in property measurement. House price indices which experience volatile swings; where the price is increasing one month, decreasing the next, and returning to an increase the month after that, should be penalised under this robustness system, with the magnitude of this volatility taken into account. A model exhibiting this type of behaviour, by definition, would be anti-autocorrelated, violating extensive existing findings (Case and Shiller, 1988; Englund and Ioannides, 1997).

Comparing these discrepancies in magnitude for successive months is similar to

Goh, Costello, and Schwann, 2012's approach of minimising discrepancies in a given month using random sampling; models which strip out more noise should result in a smoother index, which more closely and accurately captures the real behaviour of the strongly cyclical property market. Thus, we will evaluate the various models explored in this chapter by looking at the absolute monthly change in the momentum of the index. For example, an index which rises by the same relative amount each month would have an average monthly change in momentum, or *smoothness* value of 0%.

Wang and Zorn, 1997 posit that an index should be defined by its use in practice, rather than being excessively influenced by the higher level concerns of extreme statistical soundness and precision. They found in their analysis that much of the disagreement and debate over the methodology of indices can be distilled to largely unrecognised contention over the actual intended application of the models. In accordance with this, our mathematical concept of index smoothness can instead by expressed as the most robust index being the one which investors would seek to hold if house price indices were an openly traded market, as per the recommendations of Englund, Hwang, and Quigley, 2002; Shiller, 2003.

Investors seek to hold a maximally diversified portfolio of assets, which minimises risk and maintains return (Maguire, Moser, et al., 2014). It has been shown that low-volatility portfolio investment produces the most reliable and consistent returns over the a long period of time, contrary to what might seem logical (Maguire, Kelly, et al., 2017). It may seem reasonable that, on average, the high risk investor, if sufficiently skilled at trading, should accumulate a higher return on capital, owing to the outsized amount of risk they take through holding a more volatile position. Despite this, the opposite effect has been statistically shown to be the reality; low volatility portfolios perform best over a long time horizon (Baker, Bradley, and Wurgler, 2011).

As such, there appears to be a strong link between portfolio diversification and low volatility. It has been theorised that the level of diversification in a portfolio can be measured by analysing how independent sources of information combine in order to smooth the volatility of the portfolio as a whole (Choueifaty, Froidure, and Reynier, 2013). Relating this to house price indices, suppose that a number

of distinct models were published openly by a variety of organisations; investors would naturally seek to hold the optimal combination of those sources, such that the long term volatility of those portfolios is minimised (Kuo and Li, 2013). In other words, they would seek to hold the smoothest composite index. In summation, we conclude that, in practice, the optimal house price index is the smoothest house price index.

3.4 Comparison of results

Table 3.2 shows the results of each stage of our proposed sparse and frugal *GeoPrice* model, alongside the CSO's raw RPPI model and the simple average and median measures. As shown in the table, our stage four house price index managed to exceed the *smoothness* value of the RPPI, achieving a result of 2.83% versus the RPPI's 3.35%. Across the entire sample period, the absolute maximum monthly percentage change was a drop of 6.7% in January 2013, versus the RPPI's largest change being a gain of 8.1% in December 2012. However, the correlation between the monthly changes of the models was relatively low, at $r = 0.43$. This would suggest that they each contribute somewhat different pieces of information.

According to Quigley, 1995, hybrid models which aggregate the results of both hedonic regression and repeat sales models tend to perform more robustly than either of these models in isolation. We found a similar phenomenon in our data; combining our frugal index with the RPPI as a composite index (with a 56.1% weight for our frugal index and 43.9% weight on the RPPI) reduced volatility and increased the smoothness metric to 2.51%.

Figure 3.6 shows the results of our frugal model, the RPPI and the composite house price index on a plot. Despite our frugal index outperforming the CSO's RPPI index in robustness, the composite index is the superior index of the three. As discussed previously, this is the index which investors would choose to hold, if both indices were available to trade on an open market.

TABLE 3.2: Statistical results for a selection of price indices

Metric	Mean (%) [a]	Median (%) [a]	Max (%)	Min (%)	St.Dev (%)	Smooth (%)
Raw average	7.01	5.74	+31.1	-17.9	9.23	12.40
Raw median	5.06	4.07	+23.8	-15.2	6.79	8.42
Stage One	3.85	3.23	+11.1	-15.6	4.81	6.37
Stage Two	2.58	2.70	+9.19	-7.38	3.31	4.47
Stage Three	2.72	2.22	+7.33	-8.38	3.38	3.76
Stage Four	2.05	1.64	+5.41	-6.67	2.55	**2.83**
RPPI	2.16	1.61	+8.06	-5.50	2.73	**3.35**

[a] These metrics are applied to the absolute monthly percentage changes of the corresponding index.

FIGURE 3.6: RPPI, *GeoPrice* and composite indices from 01-2010 to 06-2015 (inclusive)

3.4.1 Advantages of the *GeoPrice* model

Contrary to the claims of O'Hanlon, 2011, the frugal data available through stamp duty returns is sufficient for constructing a house price index which not only matches, but exceeds that of the CSO's RPPI model. Admittedly, while the *GeoPrice* index does not offer a vast improvement in smoothness over the existing RPPI, it does have a major advantage in that it only makes use of the publicly available sale price, date of sale and address, rather than requiring the large selection of property attributes which the CSO must collect privately in order to produce the RPPI index.

This makes the proposed model highly flexible and easy to deploy. It is possible for the algorithm to operate on any dataset which consists of no more than the three attributes mentioned previously. It is constructed in such a way that it automatically controls for noise, property mix bias and outlier data. The algorithm can be re-computed at will, as soon as new data becomes available. As a result, the sparse and frugal model can produce results in a matter of days following new data being collected, rather than a matter of months, as is the case with most hedonic regression models, as discussed previously.

Furthermore, the *GeoPrice* index could be expanded through application to asking prices for properties which have not yet been sold. This would potentially result in a forecasting effect, whereby the changes in the coming three-to-six months could be estimated, simply by using data collected from a property listing portal. An index

could also be produced on any sub-region of the dataset, provided there is an ample number of transactions to generate a reliable signal. These expanded use cases of the algorithm are trivial to implement, once the data is available. We will explore further applications and enhancements to the algorithm later in this thesis.

3.4.2 Limitations of the *GeoPrice* model

One of the main limitations of this sparse and frugal model is the time taken to compute the index. Due to the heavy use of nearest neighbour searches across the dataset, computation takes a significant amount of time (circa 14 hours on the dataset presented in this thesis). While this doesn't prevent frequent and rapid computation of the index once new data is released, it does increase the difficult of computing many variations of the index.

Furthermore, the nearest neighbour searches in stage four of the algorithm are of quadratic complexity. This means that the execution time will increase quadratically as more properties are added, limiting the expansion to much larger datasets. Methods of addressing and improving upon this drawback will be explored in Chapter 4, which will be intended to make the *GeoPrice* model more scalable and performant.

3.5 Chapter Summary

Timely measurement of property market trends is of high importance to market stakeholders, particularly policy makers, and is of critical importance to understanding the behaviour of the housing market. Empirical evidence on the matter has theorised that introducing real estate derivative markets would bring large economic benefits and aid in rapidly adjusting the valuation of properties towards an equilibrium between supply and demand (Englund, Hwang, and Quigley, 2002). It has also been suggested that these changes would result in lower rents and more timid market movements owing to speculative investment (Iacoviello and Ortalo-Magne, 2003; Quigley, 1999). Our sparse and frugal method makes strides in supporting this product, by dramatically reducing the publication lag on housing market statistical measures. Further progress on this front will be made in the coming chapters of this thesis.

Critics of this model may argue that, over a long period of time, precisely calibrated statistical models provide a more transparent view of gradual changes in the property market. This may well be true, however, it could also be argued that the primary function of a house price index, or any model for that matter, is to communicate immediate and timely changes in the instrument it is intended to measure. According to Wang and Zorn, 1997, there is little use in striving to achieve statistical accuracy or perfection if that goal does not translate to practical use, improved decision making and better economic results.

Governments and policy makers will typically make decisions on the short-term movements of key economic indicators, as their reaction function needs to be timely in order to avoid any troublesome situation spiraling out of control. As such, housing market participants typically look to frequently updating, active measures, as discussed in detail in Section 2.4.

The initial formulation and application of the *GeoPrice* index is a proof of concept algorithm, built on public, parsimonious data with an accessible methodology, aimed at filling this niche. The model, in its current state, has already been demonstrated to be capable of outperforming, albeit mildly, the accuracy of a conventional hedonic regression model, despite a vastly more sparse set of input data. Furthermore, this outperformance was achieved using a dataset which is entirely public, leading to greater transparency and reproducibility of the index.

In future chapters, improvements to the efficiency, smoothness and performance of the model will be explored, aiming to boost the accuracy and scalability of the methodology. Applications of the model to datasets which are both larger and distinct in nature will be presented, in order to demonstrate the suitability of the model to wider use.

Chapter 4

GeoTree: A data structure for $O(1)$ geospatial search, enabling a real-time property index [1]

The *GeoPrice* model presented in Chapter 3 demonstrated the potential of leveraging geospatial auto-correlation to construct a smooth, accurate house price index model capable of outperforming a hedonic regression model, without the need to acquire a rich set of descriptive attribute data on each property. Through the use of geospatial matching with similar neighbours, the *GeoPrice* index demonstrated that the likelihood of proximate properties sharing similar price-influencing characteristics was enough to generate relevant comparables across different time periods, from which a price index could then be generated.

One of the major drawbacks of this initial formulation of the *GeoPrice* index, as discussed in Section 3.4.2, is the slow execution time of the computationally-demanding methodology. Since the model is required to search for neighbours of each property transacted in month X in every month prior to month X, an intensive, nested nearest neighbour search must be performed across a significant number of data points.

In order to improve upon this limitation of the model, this chapter will focus on

[1] This chapter is adapted from *GeoTree: A Data Structure for Constant Time Geospatial Search Enabling a Real-Time Property Index* (Miller and Maguire, 2021), which was expanded upon in the consolidated journal article *A real-time mix-adjusted median property price index enabled by an efficient nearest neighbour approximation data structure* (Miller and Maguire, 2022).

introducing a more efficient method of determining neighbours of each given property. A novel data structure, the *GeoTree*, will be outlined and applied to the dataset used in Chapter 3, demonstrating a performance improvement of multiple orders of magnitude. This result is achieved by replacing each $O(n)$ neighbour search with an $O(1)$ search. The cost of this efficiency gain is a slight reduction in accuracy, due to limitations of the encoding methodology, however, we will demonstrate that the impact of this on the *GeoPrice* index is minor.

4.1 Complexity of high-volume geospatial search queries

Large scale datasets are a hot topic in computer science. Each one tends to present its own problems and intricacies (Hand, 2013). The Nearest Neighbour (*NN*) problem is a well known and vital facet of many data mining research topics. This involves finding the nearest data point to a given point under some metric which measures the *distance* between data points. In the context of geospatial data, the *NN* problem often emerges in the form of geographical proximity search (Roussopoulos, Kelley, and Vincent, 1995).

Real world geographic data is usually represented by a pair of GPS co-ordinates, which pinpoint any location on Earth with unlimited precision. As a result of their structure, computing the distance between pairs of points in order to find the *nearest neighbour* can be extremely slow on large datasets, given the quadratic nature of applying the naive nearest neighbour algorithm (Ramírez-Gallego, Krawczyk, et al., 2017; Zhang, Mamoulis, et al., 2004).

This impediment often requires further expansion to finding the k nearest neighbours (k-NN), or all neighbours within a certain range, which increases the computational complexity through requiring a sorting of the distance matrix, in order to extract a ranking of points by proximity. It is extremely computationally expensive to compute and rank these distances on large datasets Safar, 2005. A computationally cheap method of solving this problem would vastly improve the scalability of proximity based algorithms Roussopoulos, Kelley, and Vincent, 1995, particularly including our sparse and frugal property price index model proposed in Chapter 3.

4.2 Methods of geospatial search

4.2.1 Naive haversine search

The distance between two pieces of geospatial data defined using the GPS co-ordinate system is computed using the *haversine* formula (Robusto, 1957). If we wish to find the closest point in a dataset to any given point in a naive fashion, we must loop over the dataset and compute the haversine distance between each point and the given, fixed point. This is an $O(n)$ computation. If the distances are to be stored for later use, this also requires $O(n)$ memory consumption. Thus, if the closest point to every point in the dataset must be found, this requires an additional nested loop over the dataset, resulting in $O(n^2)$ memory and time complexity overall (assuming the distance matrix is stored). If such a computation is applied to a large dataset, such as the 147,635 property transactions used in the house price index in Chapter 3, an $O(n^2)$ algorithm can run extremely slowly even on powerful modern machines.

As GPS co-ordinates are multi-dimensional objects, it is difficult to prune and cut data from the search space without performing the haversine computation. With a considerable portion of big data being geospatial in nature, geospatial algorithms and data structures are coming under increased research attention, with the amount of personal location data available growing by approximately 20% year-on-year according to the *McKinsey Global Institute* (Lee and Kang, 2015). As such, exploring alternative methods of representing GPS co-ordinates is necessary to make algorithmic improvements and advance the feasibility of employing geospatial data in models and analysis.

4.2.2 GeoHashing

A geohash is a string encoding for GPS co-ordinates, allowing co-ordinate pairs to be represented by a single string of characters. The publicly-released encoding method was invented by Niemeyer in 2008 (Niemeyer, 2008). The algorithm works by assigning a geohash string to a square area on the earth, usually referred to as a *bucket*. Every GPS co-ordinate which falls inside that bucket will be assigned that geohash. The number of characters in a geohash is user-specified and determines the size of the bucket. The more characters in the geohash, the smaller the bucket becomes,

and the greater precision the geohash can resolve to. While geohashes thus do not represent points on the globe, as there is no limit to the number of characters in a geohash, they can represent an arbitrarily small square on the globe and thus can be reduced to an exact point for practical purposes. Figure 4.1 demonstrates parts of the geohash grid on a section of map.

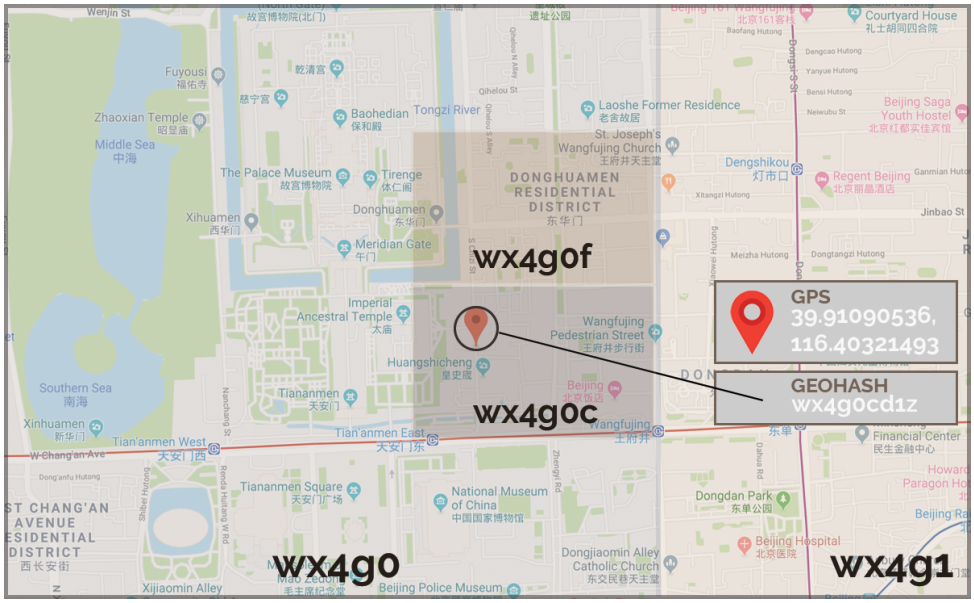


FIGURE 4.1: GeoHash algorithm applied to a map

Geohashes are constructed in such a way that their string similarity signifies something about their proximity on the globe. Take the longest sequential substring of identical characters possible from two geohashes (starting at the first character of each geohash) and call this string x. Then x itself is a geohash (ie. a bucket) with a certain area. The longer the length of x, the smaller the area of this bucket. Thus x gives an upper bound on the distance between the points. We will refer to this substring as the *smallest common bucket* (SCB) of a pair of geohashes. We define the length of the SCB as the length of the substring defining it. This definition can additionally be generalised to a set of geohashes of any size. Furthermore, we define the SCB of a single geohash g to be the set of all geohashes in the dataset which have g as a prefix. We can immediately assert an upper bound of 123,264m for the distance between the geohashes in Figure 4.2, as per the table of upper bounds in the *pygeohash* package (McGinnis, 2017).

$$\text{geohash 1: } \underbrace{c_1c_2c_3}_{\text{SCB}} x_4 \dots x_n$$

$$\text{geohash 2: } \underbrace{c_1c_2c_3}_{\text{SCB}} y_4 \dots y_n$$

$$\text{where: } x_i \neq y_i \forall i \in \{4 \dots n\}$$

FIGURE 4.2: GeoHash precision example

4.2.3 Tree structures

Geohashing algorithms have, over time, improved in efficiency and have been put to use in a wide variety of applications and research contexts (Moussalli, Srivatsa, and Asaad, 2015; Moussalli, Asaad, and Srivatsa, 2015). As stated by Roussopoulos, Kelley, and Vincent, 1995, the efficient execution of nearest neighbour computations requires the use of niche spatial data structures which are constructed with the proximity of the data points being a key consideration.

The method proposed by Roussopoulos, Kelley, and Vincent, 1995 makes use of $R-trees$, a data structure very similar in nature to the geohash (Guttman, 1984). They propose an efficient algorithm for the precise NN computation of a spatial point, and extend this to identify the exact k-nearest neighbours using a sub-tree traversal algorithm which demonstrates improved efficiency over the naive search algorithm.

A comparison of some data structures for spatial searching and indexing was carried out by Kothuri, Ravada, and Abugov, 2002, with a specific focus on comparison between the aforementioned $R-trees$ and $Quadtrees$, including application to large real-world GIS datasets. The results indicate that the $Quadtree$ is superior to the $R-tree$ in terms of insertion time due to an expensive clustering technique underpinning the methodology of the latter. As a trade-off, the $R-tree$ has faster query time. Both of these trees are designed to query for a very precise, user-defined area of geospatial data. As a result they are still quite slow when making a very large number of queries to the tree.

Beygelzimer, Kakade, and Langford, 2006 introduce another geospatial data structure, the cover tree. Here, each level of the tree acts as a "cover" for the level directly

beneath it, which allows narrowing of the nearest neighbour search space to logarithmic time in n, which would result in $O(n.\log n)$ complexity to find the nearest neighbour of every point over the entire dataset; a significant improvement over the $O(n^2)$ complexity of the naive search algorithm.

Research has also been carried out in reducing the searching overhead when the exact k-NN results are not required, and a spatial region around each of the nearest neighbours is sufficient for the use case. For example, Arya, Mount, et al., 1998 introduced an approximate k-NN algorithm called the $kd-tree$, which is a popular variant of the $R-tree$ algorithm, reducing the time complexity to $O(kd\log n)$ per query, for any given value of k. This trades off some precision in order to deliver a substantial boost in time performance.

It is often the case that ranged neighbour queries are performed as traditional k-NN queries repeated multiple times, which results in a large execution time overhead (Bao, Chow, et al., 2010). This is an inefficient method, as the lack of pinpoint precision required in a ranged query can be exploited in order to optimise the search process and increase performance and efficiency. This is a concept we leverage when constructing the *GeoTree*, in order to increase the scalibility of our frugal algorithm.

Muja and Lowe, 2014 provide a detailed overview of more recently proposed data structures such as partitioning trees, hashing based *NN* structures and graph based *NN* structures designed to enable efficient k-NN search algorithms. Again, they re-iterated the findings that *'exact search is too costly for many applications, so this has generalised interest in approximate nearest-neighbor search algorithms'*. The selection of approximation methods they investigated offered a significant speed up, however, they were largely focused on building a number of randomised trees, which resulted in substantial memory consumption in large datasets.

The *suffix-tree*, a data structure which is designed to rapidly identify substrings in a string, has also had many incarnations and variations in the literature (Apostolico, Crochemore, et al., 2016). The *GeoTree* follows a somewhat similar conceptual idea and applies it to geohashes, allowing very rapid identification of groups of geohashes with shared prefixes.

The problem can often extend to performing clustering in higher dimensional spaces. Guo, Tierney, and Gao, 2021 have developed an algorithm which can filter

out a large number of data points to deliver an implementation of *sparse subspace scattering* which maintains near identical performance, with a reduction to linear time and memory complexity on the nearest neighbour search.

The common theme within this existing body of work is the sentiment that methods of speeding up k-NN search, particularly upon data of a geospatial nature, require specialised data structures designed specifically for the purpose of proximity searching (Roussopoulos, Kelley, and Vincent, 1995). As shown, these algorithms vary in the level of speed up offered relative to their desired accuracy trade-off and memory limits, according to the specific use case motivating their development. In a similar vein, we will introduce a novel, approximate neighbourhood matching data structure, which will allow us to quickly retrieve the approximate neighbours of any property in our dataset.

4.3 The *GeoTree* data structure

The goal of our data structure is to allow efficient approximate ranged proximity search over a set of geohashes. For example, given a database of house data, we wish to retrieve a collection of houses in a small radius around each house without having to iterate over the entire database. In more general terms, we wish to pool all other strings in a dataset which have a maximal length SCB with respect to any given string. Being able to perform these searches efficiently will offer a dramatic speed up and improvement in scalability in stage two, stage three and stage four of our sparse and frugal model, detailed in Algorithm 2 and Algorithm 3.

4.3.1 High-level description

A *GeoTree* is a general tree (a tree which has an arbitrary number of children at each node) with an immutable fixed height h set by the user upon creation. Each level of the tree represents a character in the geohash, with the exception of level zero - the root node. For example, at level one, the tree contains a node for every character that occurs among the first characters of each geohash in the database. For each node in the first level, that node will contain children corresponding to each possible character present in the second position of every geohash string in the dataset sharing

the same first character as represented by the parent node. The same principle applies from level three to level h of the *GeoTree*, using the third to h^{th} characters of the geohash respectively.

At any node, we refer to the path to that node in the tree as the *substring* of that node, and represent it by the string where the i^{th} character corresponds to the letter associated with the node in the path at depth i.

The general structure of a *GeoTree* is demonstrated in Figure 4.3a. As can be seen, the first level of the tree has a node for each possible letter in the alphabet. Only characters which are actually present in the first letters of the geohashes in our dataset will receive nodes in the constructed tree. We, however, include all characters in this diagram for clarity. In the second level, the a node also has a child for each possible letter. This same principle applies to the other levels of the tree. Formally, at the i^{th} level, each node has a child for each of the characters present among the $(i+1)^{th}$ position of the geohash strings which are in the SCB of the current substring of that node. A worked example of a constructed *GeoTree* follows in Figure 4.3b.

Consider the following set of geohashes which has been created for the purpose of demonstration:

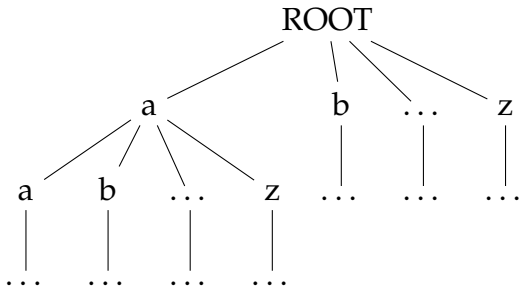
$$\{gc7j98, gc7j98, gd7j98, ac7j98, gc9aaj, gc7j9d, ac7j98, gd7jya, gc9aaj\}$$

The *GeoTree* generated by the insertion of the geohashes above with a fixed height of six would appear as seen in Figure 4.3b.

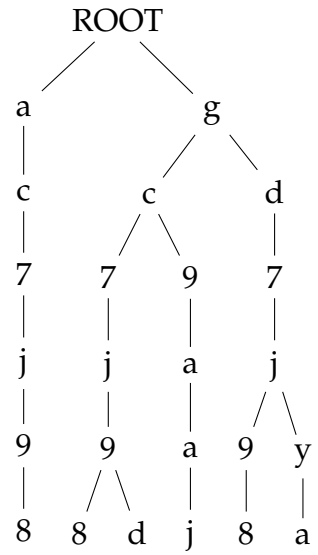
4.3.2 Data nodes

The data attributes associated with a particular geohash are added as a child of the leaf node of the substring corresponding to that geohash in the tree, as shown in Figure 4.4. In the case where one geohash is associated with multiple data entries, each data entry will have its own node as a child of the geohash substring, as demonstrated in the diagram.

It is now possible to collect all data entries in the SCB of a particular geohash substring without iterating over the entire dataset. Given a particular geohash in the tree, we can move any number of levels up the tree from that geohash's leaf



(A) *GeoTree* General Structure



(B) *GeoTree* Structure via Example

FIGURE 4.3: *GeoTree* Structure Diagrams

nodes and explore all nearby data entries by traversing the sub-tree given by taking that node as the root. Thus, to compute the set of geohashes with an SCB of length m or greater with respect to the particular geohash in question, we need only explore the sub-tree at level m along the path corresponding to that particular geohash. Despite this improvement, we wish to remove the process of traversing the sub-tree altogether.

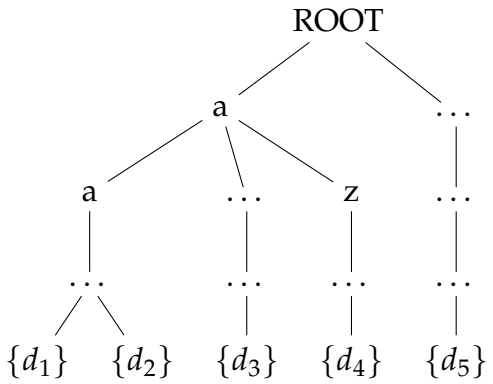


FIGURE 4.4: *GeoTree* Structure with Data Nodes

4.3.3 sub-tree caching of data nodes

In order to eliminate traversal of the sub-tree we must cache all data entries in the sub-tree at each level. To cache the sub-tree traversal, each non-leaf node receives an additional child node which we will refer to as the *list* (*ls*) node. The list node holds references to every data entry that has a leaf node within the same sub-tree as the list node itself. As a result, the list node offers an instant enumeration of every leaf node within the sub-tree structure in which it sits, removing the need to traverse the sub-tree and collect the data at the leaf nodes. The structure of the tree with list nodes added is demonstrated in Figure 4.5 (some nodes and list nodes are omitted for the sake of brevity and clarity).

4.3.4 Retrieval of sub-tree data

Given any geohash, we can query the tree for a set of nearby neighbouring geohashes by traversing down the *GeoTree* along some substring of that geohash. A longer length substring will correspond to a smaller radius in which neighbours will be returned. When the desired level is reached, the cached list node at that level

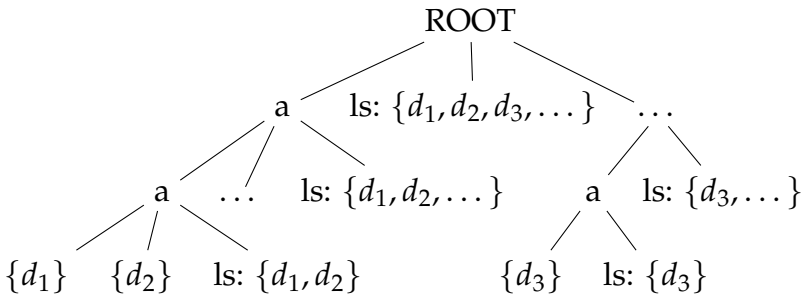


FIGURE 4.5: *GeoTree* Structure with List Nodes

can be queried for instant retrieval of the set of approximate k-NN of the geohash in question.

As a result of this design, the *GeoTree* does not produce a distance measure for the items in the *GeoTree*. Rather, it clusters groups of nearby data points and thus could be considered to be a performant geospatial clustering data structure, rather than a distance search algorithm, per se. Thus, while this does not allow for fine tuning of the search radius, it enables for a set of data points which are in near proximity to a specified geohash to be retrieved in constant time; which is our primary concern for our particular use case of geospatial property stratification.

4.3.5 Time complexity

Building (Insertion)

As hash maps offer $O(1)$ insertion, insertion of data at each level of the *GeoTree* is $O(1)$. Furthermore, insertion to the tree as a whole will take a total of h operations, where h is the height of the tree. This height is constant and fixed at creation time, thus insertion of entries to the *GeoTree* is an $O(h)$ operation, resulting in no dependence on n, the number of points in the dataset.

SCB Lookup

The $O(1)$ lookup of hash maps also means that the tree can be traversed in steps of $O(1)$ time. As the *list* nodes hold the SCB of every geohash substring possible from those in the dataset, and a maximum of h SCBs will need to be queried, it follows that any SCB lookup is at worst $O(h)$, but will generally be less than this, assuming

the user wants to perform a query for neighbours. Again, the complexity avoids any dependence on n.

4.3.6 Space complexity

As each geohash is associated with only one character at each level of the *GeoTree*, only one node on each level will hold that geohash's data entry in its list node. Thus, each data entry is inserted into one single list node at every level of the tree. Given a tree of height h, this means that the data will be stored in h different list nodes in addition to the one leaf node which the data receives. If the dataset is of size n, then there will be $(h+1) * n$ data entries stored in the tree. Thus, the overall memory requirement of the *GeoTree* is $O(hn)$ in a naive implementation. However, this can be further improved to $O(n)$ complexity by collecting a set of the data once in memory and filling the list nodes with a list of pointers to the data entries, rather than duplicating them.

4.4 Comparison with the prefix tree (trie)

The *GeoTree* data structure shares a number of similarities with the prefix tree or *trie* data structure De La Briandais, 1959. A trie is a search tree which utilises its ordering and structure to increase searching efficiency across its inserted strings. Each branch represents a character and thus as you traverse down the trie, you build the prefix of a word, working toward an entire word at each leaf node.

This is very similar to the *GeoTree*, as the geohash encodings of properties take the place of words in this use case and traversing the *GeoTree* builds prefixes of geohash strings. Both data structures make use of structure to make search more efficient, however, in the case of the *GeoTree*, the ordering has geographical significance rather than the semantic meaning in the prefix tree.

One key difference between tries and the *GeoTree* lies in the sub-tree data caching step. As the *GeoTree* relies on being able to query every entry in the sub-tree of a particular node, the caching is necessary to quickly return a large number of property records. In the case of a prefix tree, it would be necessary to enumerate every path in the sub-tree to retrieve all of the words. In the use case which the *GeoTree* is being

applied to, this would result in a significant increase in execution time over a very large dataset.

The *GeoTree* data structure could be thought of as a variant or augmentation of the trie, one which is specifically designed to give a fast, approximate solution to k-NN on geospatial datasets.

4.5 Comparison with the set enumeration tree (SE-tree)

The *SE-tree*, or *set enumeration tree*, is a power set data structure which creates a branching tree of all possible subsets of a set of variables Rymon, 1992. While this does share some basic similarities with the design of the *GeoTree*, some fundamental differences between the data structures exist. The *set enumeration tree* is a structure defined on sets which, by definition, do not consider the ordering of variables. While the SE-tree contains all possible subsets of a set of variables, it does not contain all possible ordered collections of those variables. For example, $\{A, B\}$ will be contained in the SE-tree of variables $\{A, B, C\}$, yet $\{B, A\}$ will not appear in the tree.

In the case of the *GeoTree*, all possible combinations of characters must be considered, as geohashes are sensitive to ordering. The geohash $gh1992a$, for example, corresponds to an entirely different geographical location than $hg1992a$, despite both containing the same characters in slightly different order. The *GeoTree* is designed to support this sensitivity to ordering, whereas the *set enumeration tree* is not. Furthermore, the *set enumeration tree* has no provision for the cached list nodes of data, which is perhaps the most crucial feature of the *GeoTree*. Although many interesting algorithms for traversing the SE-tree are explored in Rymon, 1992, they are irrelevant to this particular application, as the data structure in question is not designed for proximity search but for the purpose of classification.

4.6 Real-world application: the *GeoPrice* index

4.6.1 Integration with the model

In order to test the performance of *GeoTree* in practice, we applied it to the mix-adjusted median *GeoPrice* model introduced in Chapter 3. As mentioned previously,

the primary limitation of the algorithm was the algorithmic complexity and brute-force nature of the geospatial search, which impinged on its scalability to larger datasets, and restricted the introduction of further parameters.

The aim is to leverage the *GeoTree* data structure to improve the execution time, scalability and robustness of the index methodology. For the purposes of algorithmic complexity calculation, we let n be the average number of house sales present in one month of the dataset, and let t be the number of months of data in the dataset.

Stage two (Section 3.2.2.2) of the **original** algorithm is executed as follows:

$\Rightarrow$ Iterate over each month, m, of the dataset (t operations)

$\Rightarrow$ Iterate over each house, h, sold during m (n operations)

$\Rightarrow$ Iterate over houses sold in m to find the nearest to h (n operations*)

Stage four (Section 3.2.2.4) of the **original** algorithm is executed as follows:

$\Rightarrow$ Iterate over each month, m, of the dataset (t operations)

$\Rightarrow$ Iterate over each house, h, sold during m (n operations)

$\Rightarrow$ Iterate over each month prior to m, m_p ($\frac{t-1}{2}$ operations[2])

$\Rightarrow$ Iterate over houses sold in m_p to find the nearest to h (n operations*)

By introducing the *GeoTree* to the algorithm, the steps which formerly required an $O(n)$ iteration over all houses in the dataset to identify the nearest house (marked by an asterisk) now become an $O(1)$ *GeoTree* ranged proximity search operation.

There is, however, a mild trade-off. Rather than returning the closest property to the house in question, the *GeoTree* structure instead returns everything in a small area around the house (formally, it returns the maximal length non-empty SCB for that house's geohash). The bucket can then be iterated over to find the true closest property, or an alternative strategy can be employed, such as taking the median price of all houses within the small area.

[2]The number of iterations will be one less than the current month's index. Given t months in total, the number of iterations will run from 0 to $t-1$ sequentially. The mean number of iterations is thus $\frac{t-1}{2}$

However, the algorithmic change of considering the median of a nearby group of homes could be considered a diversification benefit, rather than a drawback, as it reduces the potential impact of outliers which happen to be the absolute nearest data record to the property undergoing proximity search.

4.6.2 Performance results

Table 4.1 compares the performance of the algorithms described previously with and without the *GeoTree* data structure (on a database of 279,474 property sale records), including both single threaded execution time and multi-threaded execution time (running across eight CPU cores) on our test machine. The results using the *GeoTree* are marked with a + symbol.

TABLE 4.1: Complexity and performance of the algorithms

Algorithm	Complexity	μ (1 core)[a]	σ[b]	μ (8 cores)[a]	σ[b]
Voting	$O\left(n^2t\right)$	233.54 seconds[c]	2.37%	46.73 seconds[c]	1.69%
Voting+	$O\left(nt\right)$	12.78 seconds[c]	1.68%	3.02 seconds[c]	0.69%
Stratify	$O\left(\frac{n^2t(t-1)}{2}\right)$	29.03 hours	2.41%	4.19 hours	1.89%
Stratify+	$O\left(\frac{nt(t-1)}{2}\right)$	~0.05 hours (163.89s)	1.71%	~0.01 hours (39.63s)	0.85%
Overall	$O\left(\frac{n^2t(t+1)}{2}\right)$	29.11 hours	2.43%	4.21 hours	1.90%
Overall+	$O\left(\frac{nt(t+1)}{2}\right)$	~0.05 hours (177.73s)	1.67%	~0.01 hours (43.71s)	0.79%

[a] Execution times reported are the mean (μ) of ten trials.
[b] Standard deviation (σ) reported as a percentage of the mean (μ).
[c] **Includes build time** for the dataset array / *GeoTree* on the dataset, as applicable.
[d] All algorithms computed using an AMD Ryzen 2700X CPU.
[e] All algorithms executed on the Irish Residential Property Price Register database of **279,474 property sale records** as of time of execution.

4.6.3 Correlation with the original model

Despite the algorithmic alteration of taking the median price of a group of geo-hashed nearest neighbours, as opposed to the nearest neighbour per se, the house

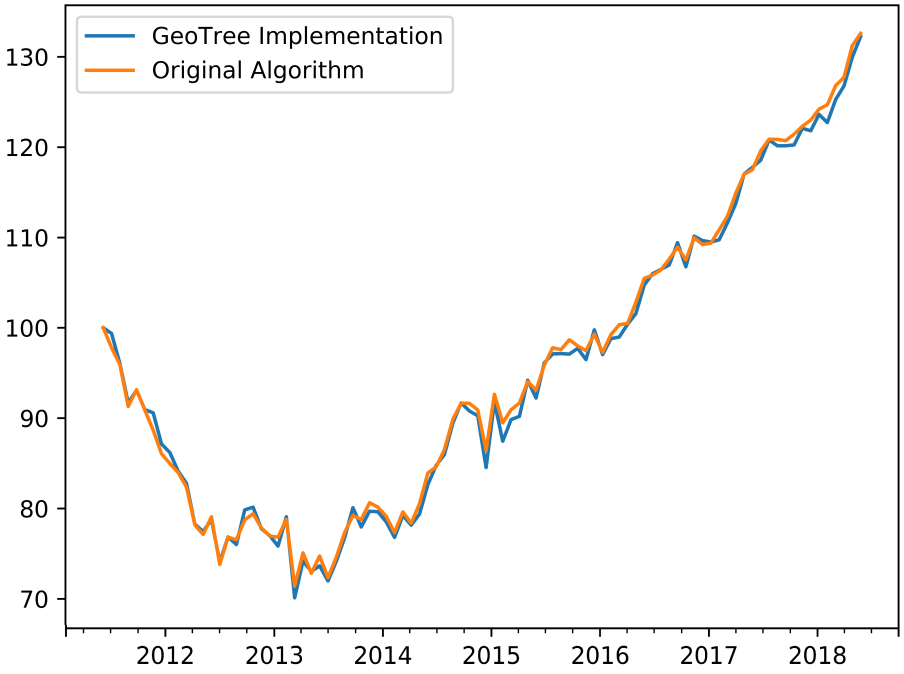


FIGURE 4.6: Sparse and Frugal House Price Index for Ireland (*GeoTree* vs Original), from 02-2011 to 09-2018

price indexes produced by the original algorithm and the *GeoTree*-enhanced version are very similar. Figure 4.6 shows both versions of the *GeoPrice* model superimposed. The two different versions yielded highly correlated outputs (Pearson's r = 0.999, Spearman's $\rho = 0.997$, Kendall's $\tau = 0.966$), revealing that *GeoTree* succeeded in delivering an almost identical index to the original, though with major performance gains in execution time.

4.6.4 Potential expansion of the model

The significant improvement to the execution time of our algorithm opens a large number of avenues for expansion of our model. Firstly, the reduction in computational complexity allows us to consider applying the algorithm to much larger datasets. As can be seen in Table 4.1, the computation time taken by the original algorithm on a relatively small dataset (Irish property sales) would balloon to a highly prohibitive runtime if applied to larger markets, such as the UK. Through utilising the *GeoTree*, this is now a feasible endeavor, given the reduction to linear-time dependence on the size of the dataset.

Other potential enhancements we could make to our algorithm include offering the ability to easily stratify a dataset beyond geospatial stratification alone. Given that an encoding of co-ordinates is being leveraged to perform the stratification, it would, in theory, be possible to encode further attributes alongside this information, if such property characteristics were available to the user. We will explore the potential of this improvement further in Chapter 5.

Furthermore, the possibility of using the index as a frequently updating source of information for a derivative market for properties, discussed in Section 3.5, is noticeably more feasible. Where the prior execution time would have been restrictive in this sense, the enhanced model allows for an index which could be recomputed on an hourly basis, for example.

4.7 Scalability testing

In order to verify the scalability of the *GeoTree*, we obtained a dataset comprising 2,857,669 property sale records for California, and evaluated both the build and query time of the data structure. Table 4.2 shows mean build time and mean query time on both 10% (∼285,000 records) and 100% (∼2.85 million records) of the dataset. In this context, query time refers to the total time to perform **100 sequential queries**, as a single query was too fast to accurately measure.

The results demonstrate that the height of the tree has a modest effect on the build time, while dataset size has a linear effect on build time, thus supporting the claimed $O(h)$ insertion per single record. Furthermore, query time is shown to remain constant regardless of both tree height and dataset size, with negligible differences in all instances, owing to the relatively small magnitude of h in any reasonable use case of the structure.

4.8 Chapter Summary

The *GeoTree* data structure is a highly performant, approximate geospatial clustering structure, which has been designed and tailored for the use case of a geospatially

TABLE 4.2: Scalability Performance of *GeoTree*

Height h	4	5	6	7	8
Build Time (10%)[a]	**17.63s** (0.08s)	**18.10s** (0.10s)	**18.46s** (0.22s)	**18.84s** (0.08s)	**19.39s** (0.09s)
Build Time (100%)[b]	**179.67s** (0.58s)	**183.80s** (0.57s)	**183.99s** (0.52s)	**192.06s** (0.60s)	**194.31s** (0.94s)
Query Time (10%)[c]	**5.1ms** (0.3ms)	**5.2ms** (0.4ms)	**5.3ms** (0.9ms)	**5.3ms** (0.4ms)	**5.3ms** (0.5ms)
Query Time (100%)[c]	**5.4ms** (1.0ms)	**5.3ms** (0.9ms)	**5.5ms** (1.0ms)	**5.7ms** (1.3ms)	**5.6ms** (1.2ms)

[a] Build Time (10%) is the total time to insert 10% of dataset (∼285,000 records)
[b] Build Time (100%) is the total time to insert 100% of dataset (∼2.85m records)
[c] Query Time consists of total time to execute 100 sequential neighbour queries on 10% and 100% of the dataset respectively
[d] Times reported are in the format $\mu(\sigma)$ calculated over ten trials

stratified property price index. The development of this data structure was necessary in order to deliver the performance-accuracy equilibrium required for our specific goals, namely; scalability, ease of use, flexibility and frequent and regular re-computation of the *GeoPrice* index.

While a small amount of resolution is lost by organising properties into the buckets required to use the *GeoTree*, it is not necessary for the *GeoPrice* model to consider the single nearest neighbour with pinpoint accuracy; the spatial auto-correlation of properties extends to neighbourhoods and housing developments, rather than just the property next door, as discussed in Section 3.2.2.2. As such, considering a basket of multiple neighbours in close proximity is likely an equivalently sound, if not superior method to our original nearest neighbour proposal, as demonstrated by the correlation statistics and visible similarity of the indices when superimposed on a chart.

The performance measures explored in this chapter demonstrate that the structure has succeeded in meeting the purposes for which it was designed. These efficiency gains will allow the *GeoPrice* algorithm to be expanded beyond the bounds

faced by the original model's methodology, both in terms of dataset scale and stratification granularity.

The application potential unlocked through lifting this algorithmic limitation will be explored in the next two chapters through exposure of the *GeoPrice* model to two new, distinct datasets. These additional use-cases are intended to demonstrate that the *GeoPrice* model, in combination to the *GeoTree* clustering data structure, can be flexibly applied to a number of different contexts, regions and dataset sizes, while maintaining reliable and consistent performance.

Chapter 5

Applying the *GeoPrice* model to listed asking prices in the Irish property market [1]

With the performance limitations of the original formulation of the *GeoPrice* index lifted through the introduction of the *GeoTree* data structure in Chapter 4, it is now possible to apply the *GeoPrice* algorithm to a larger dataset. The property dataset analysed in this chapter pertains to the same property market as the *Property Price Register* used in Chapter 3, however, instead of transacted property sales, the model will be fit on listed asking prices.

Furthermore, algorithmic enhancements will be introduced in this chapter in order to meet the flexibility goals of the model discussed in List 1. These developments will allow the *GeoPrice* algorithm to incorporate additional property attributes into the geospatial matching process, if and when they are available to the user, with the goal of further boosting the smoothness and accuracy of the model. Owing to the tailored nature of the *GeoTree*'s design, integrating these additional characteristics is straightforward and will not impact the performance nor complexity of the model.

[1] This chapter is adapted from *A rapidly updating stratified mix-adjusted median property price index model* (Miller and Maguire, 2020), which was expanded upon in the consolidated journal article *A real-time mix-adjusted median property price index enabled by an efficient nearest neighbour approximation data structure* (Miller and Maguire, 2022).

5.1 Introducing a new dataset: MyHome

MyHome Ltd., 2024 are a major player in property sale listings in Ireland. With data on property asking prices being collected since 2011, MyHome have a rich database of detailed data regarding houses which have been listed for sale. They have provided access to their dataset for the purposes of conducting this research.

5.1.1 Specification

The data provided by MyHome includes verified GPS co-ordinates, the number of bedrooms, the type of dwelling and further information for most of its listings. It consists of a total of 718,351 property listing records over the period February 2011 to March 2019 (inclusive). This results in 7,330 mean listings per month (with a standard deviation of 1,689), however, this raw data requires some filtering for errors and outliers.

It is important to note, however, that this dataset consists of asking prices, rather than the sale prices featured in the less detailed Irish Property Price Register Data, used in the initial incarnation of our house price index Section 3.2.

The study of modelling house prices based on asking prices, rather than on sale prices, has rarely been studied in the literature (Falzon and Lanzon, 2013). According to Scatigna, Szemere, and Tsatsaronis, 2014, this is primarily due to systematic differences between actual transacted sale prices and listed asking prices, owing to the fact the houses can be listed at a significantly inflated rate, or fail to sell entirely. However, as discussed in Section 3.3, Wang and Zorn, 1997 state that an index should be defined around an intended practical application and usefulness to market observers, rather than by overly stringent concerns of pinpoint statistical perfection.

As such, Falzon and Lanzon, 2013 found considerable explanatory power in the use of asking prices to model the housing market. Henneberry, 1998 also raised the issue of systematic discrepancies between transacted prices and asking prices, yet ultimately concluded that asking prices still offered statistically significant insight on the market and held a number of advantages over sale transactions; namely the inclusion of additional property characteristics in the former.

Furthermore, one could theorise that the lag between a property being listed and eventually being sold may give some forecasting power to asking prices, versus traditional house price indices which use transaction data. Indeed, Anenberg and Laufer, 2017 found that listing prices in the US forecasted the *Case-Shiller* repeat sales index a number of months in advance and outperformed house price forecasting methods which did not make use of asking price data.

5.1.2 Filtration of data

As with the majority of human collected data, some pruning must be done to the MyHome dataset in order to remove outliers and erroneous data. Firstly, not all transactions in the dataset include verified GPS co-ordinates or include data on the number of bedrooms. These records will be instantly discarded for the purpose of the enhanced version of the algorithm. They account for 16.5% of the dataset. Furthermore, any property listed with greater than six bedrooms will not be considered. These properties are not representative of a standard house on the market as the number of such listings amounts to just 1% of the entire dataset.

Any data entries which do not include an asking price cannot be used for house price index calculation and must be excluded. Such records amount to 3.6% of the dataset. Additionally, asking price records which have a price of less than €10,000 or more than €1,000,000 are also excluded, as these generally consist of data entry errors (e.g. wrong number of zeroes in user-entered asking price), abandoned or dilapidated properties in listings below the lower bound and mansions or commercial property in the entries exceeding the upper bound. Properties which meet these exclusion criteria based on their price amount to only 2% of the dataset and thus are not representative of the market overall.

In summation, 77% of the dataset survives the pruning process. This leaves us with 5,646 filtered mean listings per month.

5.1.3 Characteristics

Prior to exploring the performance of our house price index on the MyHome asking price dataset, it is useful to survey some of the fundamental characteristics of the

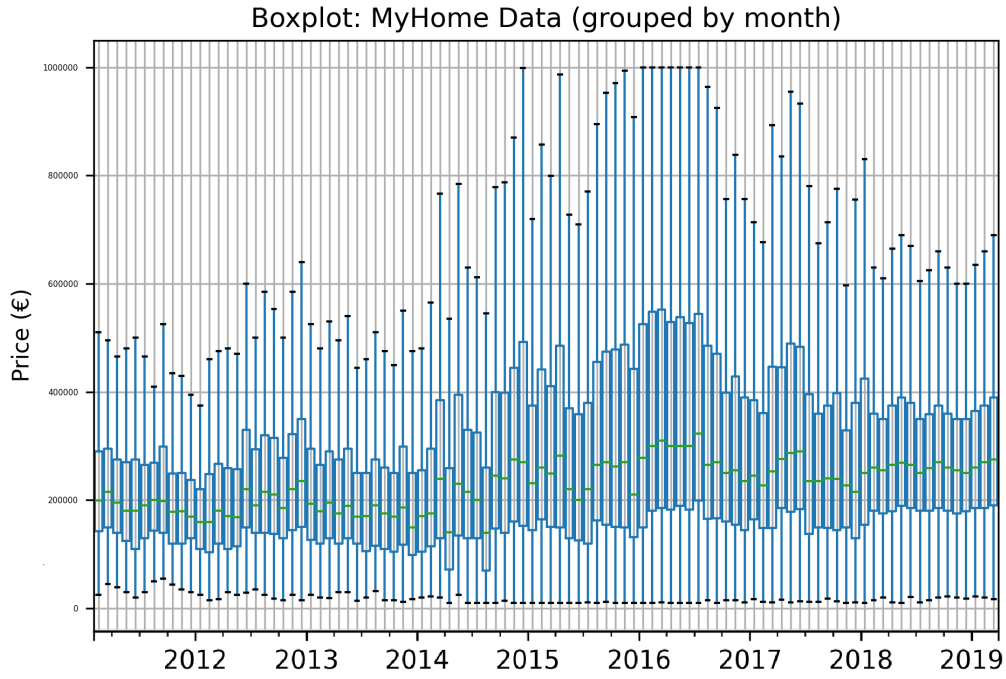


FIGURE 5.1: MyHome Listing Price Distribution from 02-2011 to 03-2019 (inclusive), grouped by month.

dataset. Figure 5.1 shows the distribution of listed prices on a month-by-month basis, across the entire dataset. It is clear that the raw asking prices are quite noisy, with the median moving around wildly from one time period to the next. The width of the interquartile range is also highly volatile, varying from a minimum range of around €125,000, to a maximum range of approximately €300,000.

It is not surprising, then, that the naive house price index consisting of simply taking the mean or the median asking price offers little-to-no insightful information about the state of the property market, as shown in Figure 5.2. The index produced by taking either of these aggregation methods frequently bounces around by high double-digit percentage swings, which is not representative of the behaviour of the market.

In terms of seasonality, the listing data does not appear to have highly influential seasonal patterns. As seen in Figure 5.3, the seasonal component is highly noisy and moves within a considerably tight range; indicative of a low impact on asking prices. This, however, seems sensible upon further thought, as the listing price is distinct from the final sale price. It is likely that the final sale price will exhibit some

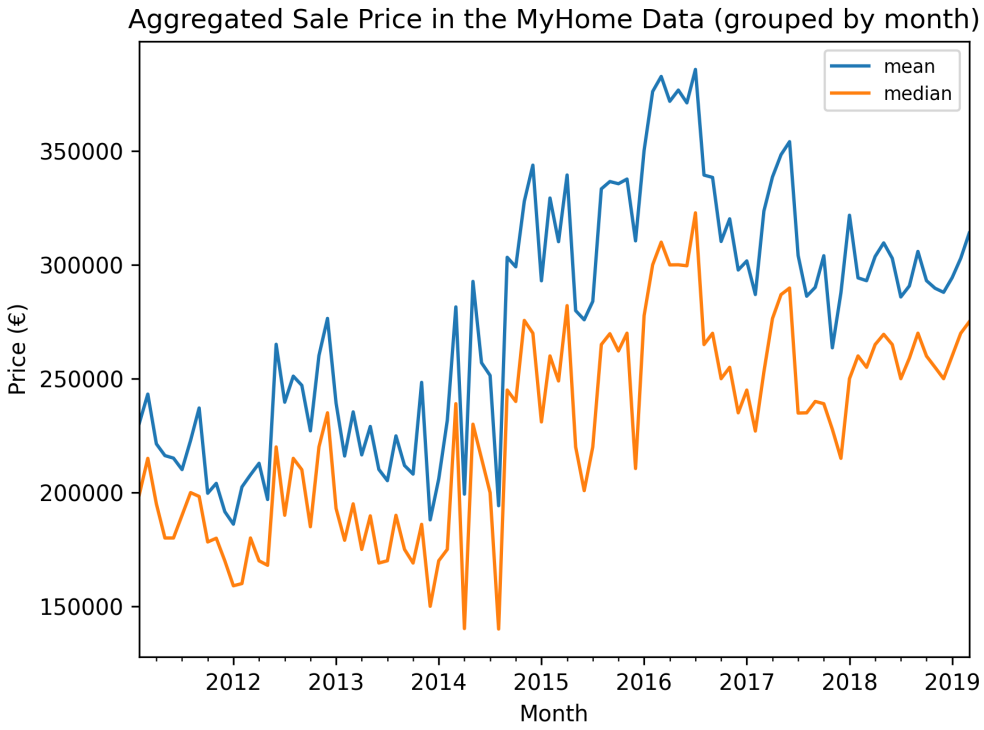


FIGURE 5.2: MyHome Mean/Median Listing Price from 02-2011 to 03-2019 (inclusive), grouped by month.

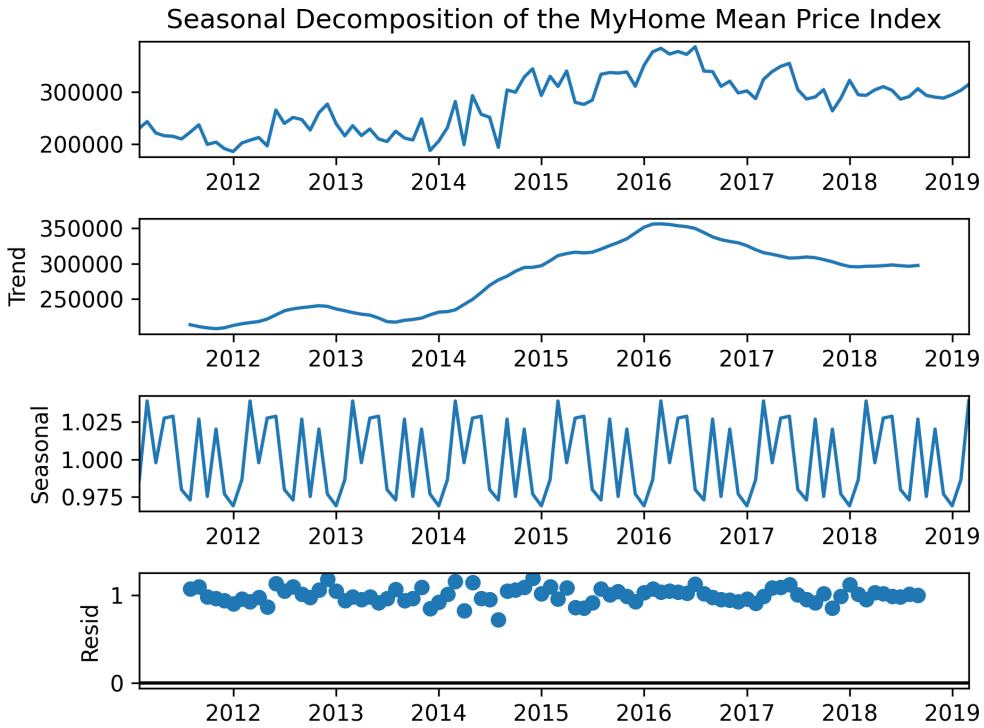


FIGURE 5.3: MyHome Mean Listing Price Seasonality from 02-2011 to 03-2019 (inclusive), grouped by month.

seasonality due to changes in demand based on the time of year, as explored in Section 3.2.1, however, a property can be listed for a considerable amount of time before a sale is completed. As the homeowner does not know in advance on which month their home will sell, it is logical then that the asking price should not contain a particularly stable or reliable seasonal signal, which would explain the volatility of the detected seasonal component.

As the MyHome dataset has been enriched with additional data on the number of bedrooms per property, we can separate the sample based on that attribute, which may reduce the volatility in looking at bucketed prices, compared to looking at the data holistically. Figure 5.4 demonstrates the distribution of the asking prices within each bucket, where buckets are based on the number of bedrooms in a given property. While this aids in curtailing the volatility to some degree, particularly in the most common categories (two and three bedrooms), it is clear that the median of each group still moves far more than what would be considered reasonable to be a reliable measure of the housing market. Even within the three bedroom bucket, for example, the interquartile range frequently exceeds a width of €250,000 in a given month.

This is reflected in Figure 5.5, which reveals that the naive median house price index still exhibits wild swings; particularly in the upper-half of the range of bedroom counts, where the range of prices in the sample is broader, as per Figure 5.4. Again, this unsophisticated price index is mostly uninformative, as the majority of information conveyed by it is simple noise, due to the variability of the basket between months.

5.1.4 Comparison with PPR dataset

The mean number of filtered monthly listings available in our dataset represents a 157% increase on the 2,200 mean monthly records used in the original algorithm's index computation (see Section 3.2.1). Furthermore, the dataset in question is significantly more precise and accurate than the PPR dataset, owing to the ability to more effectively prune the dataset. The PPR dataset consists of address data entered by hand from written documents and does not use the Irish postcode system, meaning that addresses are often vague or ambiguous. This results in some erroneous

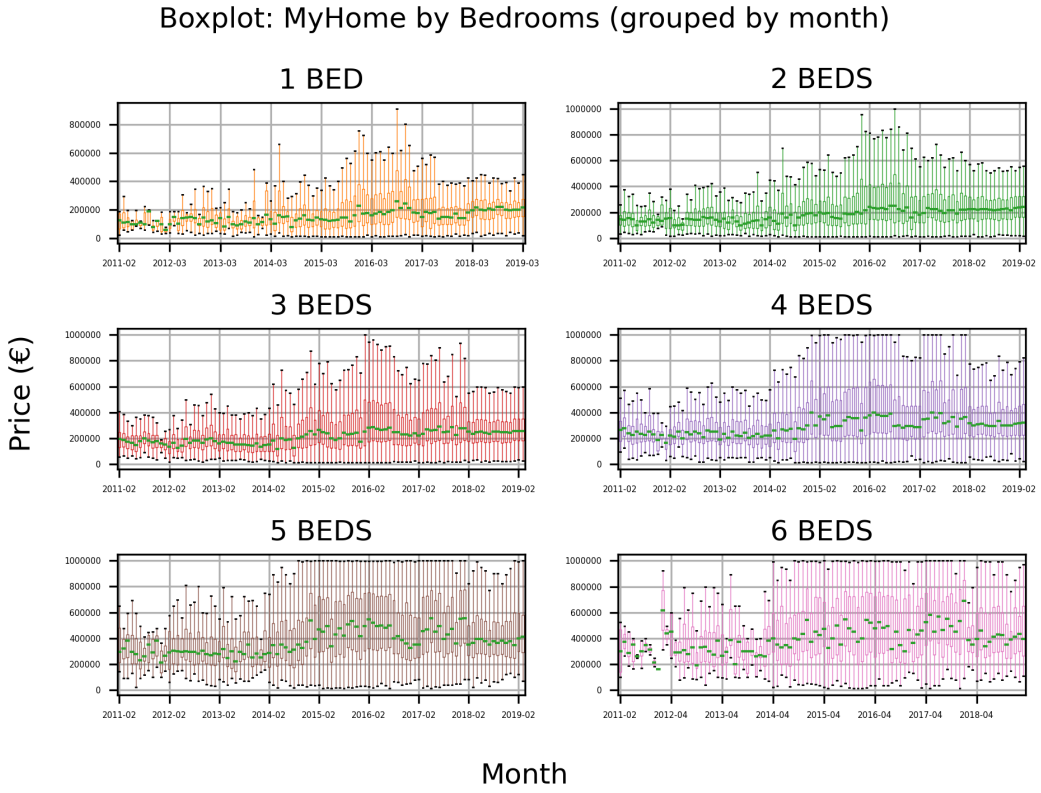


FIGURE 5.4: MyHome Listing Price Distribution per bedroom cluster from 02-2011 to 03-2019 (inclusive), grouped by month.

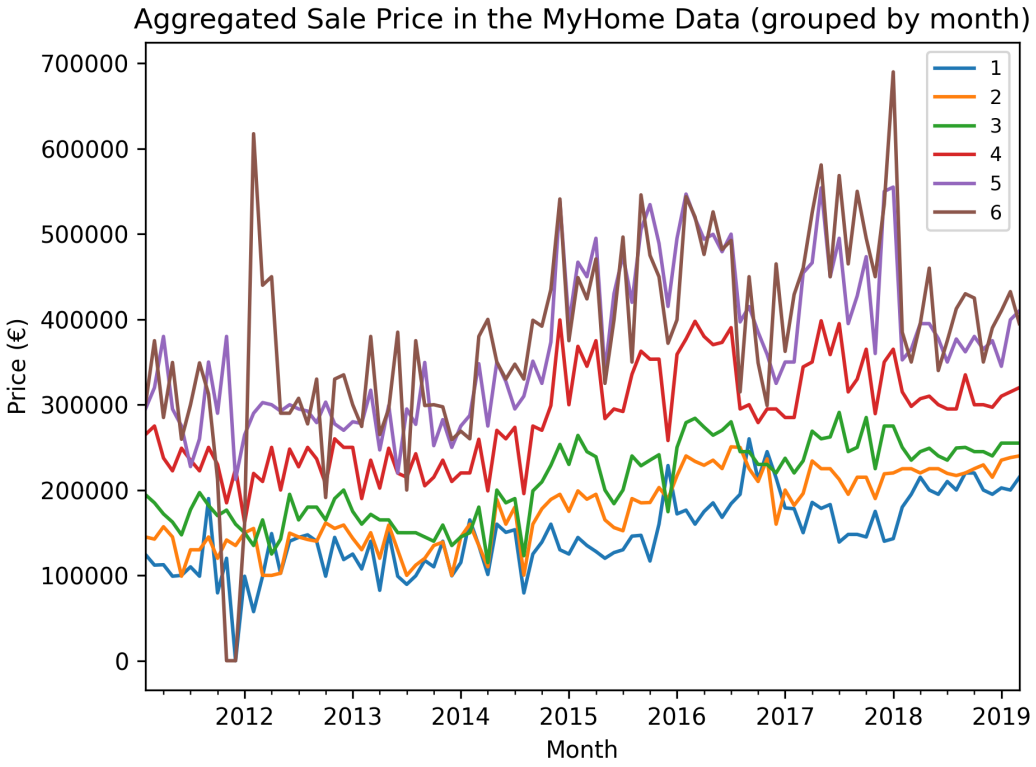


FIGURE 5.5: MyHome Median Listing Price per bedroom cluster from 02-2011 to 03-2019 (inclusive), grouped by month.

data being factored into the model computation as there is no effective way to prune this data, as discussed in Section 3.2.1.1. The MyHome dataset has been filtered to include verified addresses only, as described previously.

The PPR dataset has no information on the number of bedrooms or any key characteristics of the property. This can result in dilapidated properties, apartment blocks, inherited properties (which have an inaccurate sale value which is used for taxation purposes) and mansions mistakenly being counted as houses (see Section 3.2.1). Our dataset consists of only single properties and the filtration process described previously greatly reduces the number of such unrepresentative samples making their way into the index calculation.

The "sparse and frugal" PPR dataset was capable of outperforming the CSO's hedonic regression model with a mix-adjusted median model, as demonstrated in Section 3.4. With the larger, richer and more well-pruned MyHome dataset, further algorithmic enhancements to our model are possible.

5.2 Performance metrics for smoothness

Property prices are generally assumed to change in a smooth, calm manner over time (Clapp, Kim, and Gelfand, 2002; McMillen, 2003). As discussed in detail in Section 3.3, the smoothest index is, in practice, the most robust index. As a result of this, smoothness is considered to be one of the strong indicators of reliability for an index. However, the 'smoothness' of a time series is not well defined nor immediately intuitive to measure mathematically, therefore we wish to set out some clearer metrics for smoothness.

5.2.1 Standard deviation

The standard deviation of the time series will offer some insight into the spread of the index around the mean index value. A high standard deviation indicates that the index changes tend to be large in magnitude. While this is useful in investigating the "calmness" of the index (how dramatic its changes tend to be), it is not a reliable smoothness measure, as it is possible to have a very smooth graph with sizeable changes.

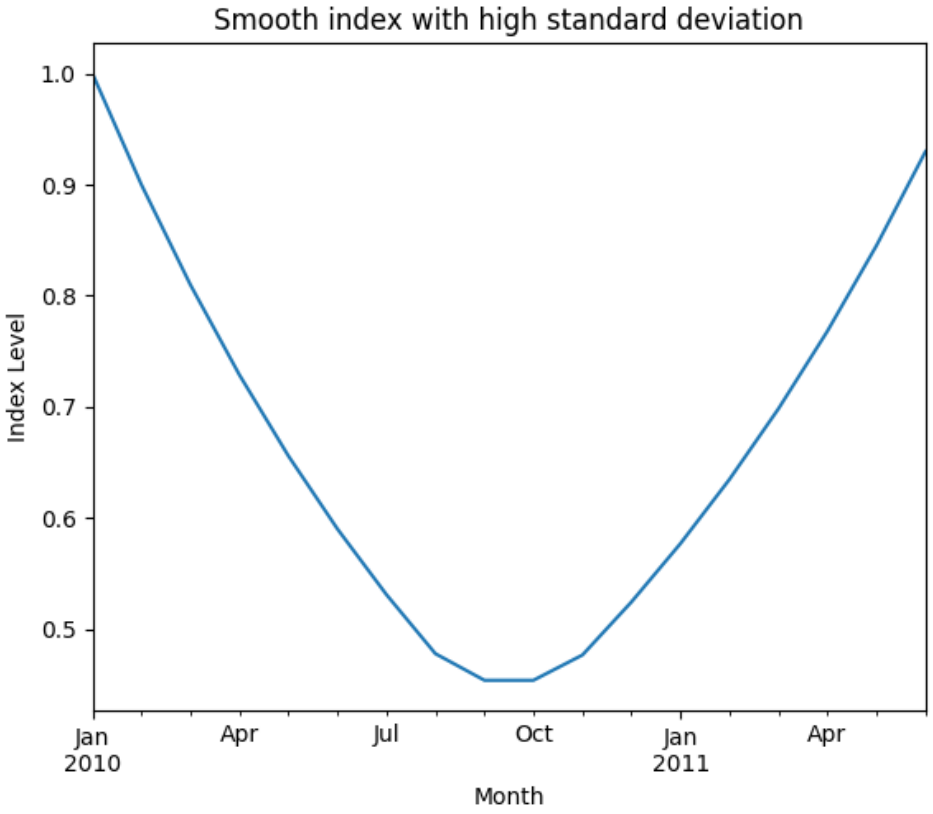


FIGURE 5.6: Sample of a smooth index with a high standard deviation

For an example, see the fictional index in Figure 5.6. In this case, the index varies very smoothly overtime, observing a downtrend with a high level of momentum (10% decrease per month), following by a bottoming (-5%, 0%, 5%), then an uptrend cycle with the same magnitude as the prior downtrend. As a result, the mean of the monthly changes is 0%, however, the standard deviation is 9.52%; almost the entire magnitude of the absolute monthly change during the up and down trends. Thus, the standard deviation is not a reliable measure for the smoothness of a time series.

5.2.2 Standard deviation of the differences

The standard deviation of the differences is a much more reliable measure of smoothness. A high standard deviation of the differences indicates that there is a high degree of variance among the differences ie. the change from point to point is unpredictable and somewhat wild. A low value for this metric would indicate that the changes in the graph behave in a more calm manner.

Referring once more to Figure 5.6, the differences of the monthly changes are now 0 in every month, except for the three months where the index is bottoming out. This results in a standard deviation of the differences score of 2.23%; which is significantly more representative of the smoothness of the sample index.

5.2.3 Mean Spike Magnitude (MSM)

Finally, we present a metric which we have defined, the *mean spike magnitude* $\mu_{\Delta X}$ (MSM) of a time series X. This is intended to measure the mean value of the contrast between changes each time the trend direction of the graph flips. In other words, it is designed to measure the average size of the 'spikes' in the graph.

Given $D_X = \{d_1, \ldots, d_n\}$ is the set of differences in the time series X, we say that the pair (d_i, d_{i+1}) is a spike if d_i and d_{i+1} have different signs. For clarity, we define zero as having a different sign to both a positive and negative number, meaning that any pair containing one single zero alongside a positive or negative number would be considered a *spike*. Then $S_i = |d_{i+1} - d_i|$ is the spike magnitude of the spike (d_i, d_{i+1}).

The *mean spike magnitude* of X is defined as:

$$\mu_{\Delta X} = \frac{1}{|S_X|} \sum_{S \in S_X} S^2$$

where:

$$S_X = \{S_1, S_2, ..., S_t\} \text{ is the set of all spike magnitudes of } X$$

Returning to our sample figure, there are two *spikes* identified in the dataset, each of size 5.0%, where the downtrend becomes flat, and where the flat index returns to an uptrend. Thus, the mean spike magnitude is simply $\frac{1}{2}.(5^2 + 5^2) = 25$, in this case. If the second spike was to have magnitude 2.5%, for example, then our mean spike magnitude would be $\frac{1}{2}.(5^2 + 2.5^2) = 15.625$. Thus, as intended, the metric measures the mean magnitude of spikes in the graph, with more penalty being applied the larger the spike. Of course, a graph could be smooth everywhere aside from one large spike, so this metric should be taken into consideration alongside other measures, such as the absolute number of spikes relative to the number of points in the dataset, or the standard deviation of the differences.

5.3 Methodological improvements

Our original central price tendency house price index was designed around a key limitation; extremely frugal data. As discussed in Chapter 3, the only data available for each property was location, sale date and sale price. The core concept of the algorithm relies on using geographical proximity in order to match similar properties historically for the purpose of comparing sale prices. While this method is likely to match certain properties inaccurately, the key concept of central price tendency is that these mismatches should average out over large datasets and cancel noise.

However, there is scope for expanding the model to include additional characteristics where available; further segmenting the geospatial strata. In order to achieve this, we must introduce a way to encode these additional attributes within a property geohash, in order to leverage the GeoTree data structure.

5.3.1 geohash$^+$

Extended geohashes, which we will refer to as geohash$^+$, are geohashes which have been modified to encode additional information regarding the property at that location. Additional parameters are encoded by adding a character in front of the geohash. The value of the character at that position corresponds to the value of the parameter which that character represents. Figure 5.7 demonstrates the structure of a geohash$^+$ with two additional parameters, p_1 and p_2.

$$\text{geohash}^+\text{: } \underbrace{p_1 p_2}_{+} \underbrace{x_1 \dots x_n}_{\text{geohash}}$$

FIGURE 5.7: geohash$^+$ format

Any number of parameters can be prepended to the geohash. In the context of properties, this includes the number of bedrooms, the number of bathrooms, an indicator of the type of property (detached house, semi-detached house, apartment etc.), a parameter representing floor size ranges and any other categorical variable desired for comparison.

Alternative applications of geohash$^+$ could extend to scenarios such as public transport mapping lookups. Say, for example, a number of public transport stations are included within a city center. Some of these may be bus stops, others train stations and a number may be underground stations. For the purpose of example, we could encode the precise location of each of these with a geohash string, based on their GPS co-ordinates. If a user was to request a list of all public transport hubs nearby, the user's live geohash location could be used with the GeoTree to rapidly return all of these locations, with the size of the common bucket adjusted based on their distance threshold preferences.

Suppose that the user then requested to filter on *only* underground stations. We could give each type of station a particular encoding; say that bus stops are encoded with B, train stations are encoded with T and underground stations are encoded with U. Then, a geohash$^+$ can be generated for each of the public transport hubs, where the encoded type of the station is prepended to the geohash. These geohash$^+$

identifier could then be used with the GeoTree data structure to return all hubs of a specific type, or types, within a close range around the user's live location.

5.3.2 GeoTree integration with geohash$^+$

Due to the design of the GeoTree data structure, a geohash$^+$ will be inserted into the tree in exactly the same manner as a regular geohash (see: Chapter 4). If the original GeoTree had a height of h for a dataset with h-length geohashes, then the GeoTree accepting that geohash extended to a geohash$^+$ with p additional parameters prepended should have a height of $h + p$. However, both of these are fixed, constant, user-specified parameters which are independent of the number of data points, and hence do not affect the constant-time performance of the GeoTree, relative to the number of dataset points.

The major benefit of this design is that the ranged proximity search will interpret the additional parameters as regular geohash characters when constructing the common buckets upon insertion, and also when finding the SCB in any search, without introducing additional performance and complexity drawbacks.

5.3.3 Enhancing our model through bedroom factoring

In order to enhance our price index model, we prepend a parameter to the geohash of each property representing the number of bedrooms present within that property. As a result, when the GeoTree is performing the SCB computation, it will now only match properties which are both nearby and share the same number of bedrooms. This allows the index model to compare the price of properties which are more similar across the time series and thus should result in a smoother, more accurate measure of the change in prices over time.

The technical implementation of this algorithmic enhancement is handled almost entirely by the GeoTree automatically, due to its design. As described previously, the GeoTree sees the additional parameter no differently to any other character in the geohash and due to its placement at the start of the geohash, the search space will be instantly narrowed to properties with matching number of bedrooms, x, by taking the x branch in the tree at the first step of traversal.

5.4 Results

We ran the algorithm on the MyHome data without factoring any additional parameters as a control step. We then created a GeoTree with geohash$^+$ entries consisting of the number of bedrooms in the house prepended to the geohash for the property.

5.4.1 Smoothness and time series comparison

Table 5.1 shows the performance metrics previously described applied to the algorithms discussed in this paper: Original PPR, PPR with GeoTree, MyHome without bedroom factoring and MyHome with bedroom factoring. While both the standard deviation of the differences and the MSM show that some smoothness is sacrificed by the GeoTree implementation of the PPR algorithm, the index running on MyHome's data without bedroom factoring approximately matches the smoothness of the original algorithm. Furthermore, when bedroom factoring is introduced, the algorithm produces by far the smoothest index, with the standard deviation of the differences being 26.2% lower than the PPR (original) algorithm presented in Maguire, Miller, et al., 2016, while the MSM sits at 58.2% lower.

If we compare the MyHome results in isolation, we can clearly observe that the addition of bedroom matching makes a very significant impact on the index performance. While the trend of each graph is observably similar, Figure 5.8 demonstrates that month to month changes are less erratic and appear less prone to large, spontaneous dips. Considering the smoothness metrics, the introduction of bedroom factoring generates a decrease of 26.8% in the standard deviation of the differences and a decrease of approximately 48.4% in the MSM. These results show a clear improvement by tightening the accuracy of property matching and are promising for the potential future inclusion of additional parameters such as bedroom matching should such data become available.

Figure 5.8 corresponds with the results of these metrics, with the *MyHome data (bedrooms factored)* index appearing the smoothest time series of the four which are compared. It is important to note that the PPR data is based upon actual sale prices, while the MyHome data is based on listed asking prices of properties which are up for sale and as such, may produce somewhat different results.

TABLE 5.1: Index Comparison Statistics

Algorithm	St. Dev	St. Dev of Differences	MSM
PPR (original)	16.524	2.191	23.30
PPR (GeoTree)	16.378	2.518	29.78
MyHome (without bedrooms)	12.898	2.209	18.91
MyHome (with bedrooms)	12.985	1.617	9.75

It is a well known fact that properties sell extremely well in spring and towards the end of the year. Furthermore, the months towards late summer, as well as the start of the year, tend to be the least busy periods in the year (Paci, Beamonte, et al., 2017). We demonstrated that these findings were observed in the Property Price Register data in Section 3.2.1.

These phenomena can be observed in Figure 5.8 where there is a dramatic increase in the listed asking prices of properties in the spring months and towards the end of each year, while the less popular months tend to experience a slump in price movement. As such, the two PPR graphs and the MyHome data (bedrooms not factored) graph are following more or less the same trend in price action and their graphs tend to meet often, however, the majority of the price action in the MyHome data graphs tends to wait for the popular selling months. The PPR graph does not experience these phenomena as selling property can be a long, protracted process and due to a myriad of factors such as price bidding, paperwork, legal hurdles, mortgage applications and delays in reporting, final sale notifications can happen outside of the time period in which the sale price is agreed between buyer and seller.

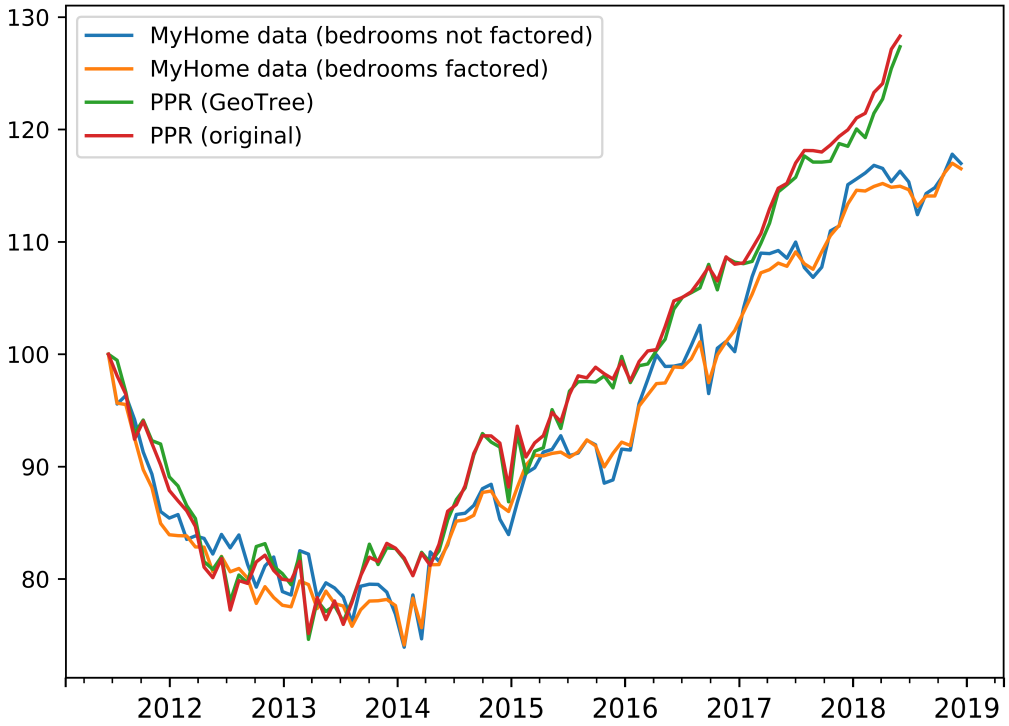


FIGURE 5.8: Comparison of index on PPR and MyHome data sets, from 02-2011 to 03-2019 [data limited to 09-2018 for PPR]

5.5 Chapter Summary

The introduction of bedroom factoring as an additional parameter in the pairing of nearby properties has been shown to have a profound impact on the smoothness of the mix-adjusted median property price index, which was already shown to outperform a popularly used implementation of the conventional hedonic regression model. This improvement was made possible due to the acquisition of a richer data set and the development of the *GeoTree* structure, which greatly increased the performance of the algorithm. There is future potential for the introduction of further property characteristics (such as property type, for example) in the proximity matching stage of the algorithm, should such data be acquired.

Furthermore, the bespoke design of the *GeoTree* data structure used ensures that minimal computational complexity is added when considering the technical implementation of this algorithmic adjustment. As a result of this, the index can be computed quickly enough that it would be possible to have real-time updates to the price index, if a sufficiently rich stream of continuous data was available to the algorithm. Large property listing websites, such as Zillow, likely have enough *live*, incoming data that such an index would be feasible to compute at this frequency, however, this volume of data is not publicly available for testing.

These enhancements and findings demonstrate that the *GeoPrice* algorithm meets the generality and flexibility goals outlined in List 1, alongside its possession of the the ability to incorporate additional information to improve modelling accuracy, where such data is available.

In the coming chapter, the *GeoPrice*'s applicability to a wide range of property modelling use-cases will be proven further through expansion of the model to a sale transaction dataset from an entirely different region, one with a much larger pool of properties and different market dynamics to those already analysed.

Chapter 6

GeoPrice: Building a property price index for the UK housing market

In the preceding chapters, the feasibility of applying the *GeoPrice* property price index model to both frugal sale transactions and listed asking prices in Ireland has been demonstrated. However, Ireland's property market is relatively small, with the total size of the housing stock estimated to have been approximately 1.86 million at most, as of 2019 (Kennedy and Myers, 2019). By contrast, the BRE reports that the total number of residential households in the UK is approximately 27.83 million, as of 2017 (Piddington, Nicol, et al., 2020).

As a result of this, the number of monthly sale transactions in the UK is much larger than that of Ireland, as will be demonstrated later in this chapter. Applying the *GeoPrice* algorithm to UK property transaction data will allow the model to benchmarked on a dataset which should be conducive to lower noise, in addition to comparing the index against a more feature-rich and robust hedonic regression model than that of the RPPI (Section 3.1): the Office for National Statistics (ONS) house price index (HPI).

The viability of applying the *GeoPrice* methodology to this larger, more expansive dataset is the result of the *GeoTree* data structure introduced in Chapter 4. The sheer volume of monthly transacted property sales in the UK would render the original index formulation (presented in Chapter 3) too slow to compute within a reasonable amount of time, defeating one of the key purposes of the model, which is to be a more timely measure of housing market trend.

Furthermore, given the ample number of monthly transactions available, the potential of generating various sub-indices for different regions in the UK by filtering the strata used in the index computation can now be explored; a key advantage of mix-adjusted median models discussed in Section 2.3. This analysis intends to demonstrate that the index is capable of meeting the final of the unfulfilled methodology goals laid out in List 1.

6.1 The ONS hedonic regression House Price Index

Prior to setting out the dataset which we will be using for our house price model, we will outline the benchmark which our index will be tested against. The Office for National Statistics produce the de-facto standard house price index in the UK jointly with HM Land Registry, Registers of Scotland and Land and Property Services Northern Ireland. This section will set out the use cases of the index, the data feeding into their model, the methodology used and the strengths and limitations of the index.

According to the ONS, house price statistics are used for a wide variety of decision-making purposes, many of which have wide-reaching impacts (ONS, 2023b). Indeed, the key users of house price indices asserted by the ONS; central government, financial institutions, local authorities, property developers and estate agents, resonates with our extensive discussion of housing market stakeholders in Section 2.4.

6.1.1 Dataset

The ONS house price index uses a variety of data sources to fit its hedonic regression model, namely: sale transaction data from the land registries of the UK; property attribute data provided by the *Valuation Office Agency*, who are responsible for organising properties into tax bands according to their value; and privately produced neighbourhood quality data, provisioned by *CACI Ltd*.

These datasets combine to give a relatively rich set of property characteristics for fitting the hedonic regression, which we will explore in more detail by outlining each of the information sources independently.

6.1.1.1 Property sale transaction data

Property sale transaction data is collected by HM Land Registry and published publicly as the Price Paid Dataset, for all properties in England and Wales [1]. The data is updated on a monthly basis; typically on the final business day of each month, with data pertaining to the month prior. Thus, there is a lag of just under one month from the end of the period in question to the publication of the sale transactions for that period.

While this dataset does not include a large number of attributes for each property, it is significantly richer than the Property Price Register dataset which we used to fit our initial sparse and frugal model in Section 3.2.1. The key attributes of interest are:

- Sale Price
- Sale Date
- Postcode
- Property Type (detached, apartment, etc.)
- New Build / Existing Build
- Assortment of Address Fields
- Private or Commercial Sale Indicator

As such, a significant amount of mix-adjustment could be carried out on this dataset alone, which we will explore later in this chapter. While this dataset covers the universe of sale transactions which occur in England & Wales, transaction data for Scotland is compiled separately by the *Registers of Scotland*, who make this data available only as a costly paid service [2]. The data attributes provided by the Registers of Scotland sale data are almost identical to the Price Paid Dataset and thus these sources are pooled by the ONS, to give a property sale transaction dataset which covers the entirety of the United Kingdom.

[1] See: https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads.
[2] See: https://www.ros.gov.uk/data-and-statistics/data-reports

6.1.1.2 Property attribute data

The primary source of attribute data for properties in England & Wales is the *Valuation Office Agency*'s (VOA) council tax valuation list. According to the ONS, this source of data contains robust attribute information on all residential properties in England & Wales. The attributes in this dataset which are considered by the ONS for the purpose of fitting their hedonic regression house price index model are:

- Number of *habitable* rooms
- Floor area, in m^2

In the VOA data, a *habitable* room is defined to be any room aside from a bathroom, conservatory, kitchen or utility room.

The alternative dataset used to supply equivalent attribute data for Scottish sale transactions is the *Energy Performance Certificate* (EPC) database. Every property sold in the United Kingdom since January 2009 requires an EPC, by law. Therefore, this dataset offers a complete set of habitable room and floor area data for sales transacted in Scotland. In the EPC dataset, a *habitable* room is defined to be a living room, sitting room, dining room, bedroom, office/study or a non-separated conservatory (ONS, 2023c).

The EPC dataset is also available in England & Wales, as an alternative to the VOA data, however, the ONS have chosen to mix the two datasets for the different regions, rather than using a single, unified data source for the entire model. The reasoning behind this decision is not communicated by the ONS in their methodology documentation. Furthermore, according to the ONS, there are significant disagreements in the data reported by each of these sources. They found that for houses, the floor area reported by the VOA data is between 2% and 10% **larger** than the EPC dataset, while in the case of apartments, the VOA data reports a floor size which is around 40% **smaller** than the EPC dataset (ONS, 2023d). These discrepancies are particularly concerning, considering that the measurement disagreement for houses and apartments have opposite signs; increasing the risk that one may be overestimated, while the other is underestimated.

In terms of matching the Price Paid Data to the Valuation Office Agency data, ONS achieve a robust match rate of around 95%, owing to the well-defined nature of addresses in England & Wales, with their mature postcode system. On the other hand, the Scottish EPC dataset does not realise quite as high a match rate with the Registers of Scotland transaction data; sitting at around 70% (ONS, 2023c). However, this should not have a significant impact on the explanatory power of the model, assuming that the missing records have no discernible patterns.

6.1.1.3 Neighbourhood quality data

The final major data input for the ONS hedonic regression model is neighbourhood quality data, which is sourced from a private company, *CACI* [3]. The *Acorn* dataset was primary designed to classify consumers into different types based on the neighbourhood they live in, in order to allow businesses to better market to the demographics and typical behaviours of each postcode.

The ONS has decided to use this dataset to capture information about the neighbourhood as a regressor in their hedonic regression model, as they posit that this should impact the value of a property (ONS, 2023a). Indeed, our discussion in Section 3.2.2.2 established grounds behind the theory of why geospatial stratification works, with some of the effect stemming from co-located houses having similar environmental factors and resident behaviours.

The hedonic regression used by the ONS includes a total of eighteen *Acorn groups* as categorical variables, with each property taking on a value of one in the corresponding Acorn group for its postcode and zero in all other groups. Figure 6.1 lists the Acorn groups and their corresponding identifiers, which feed into the house price index.

One potential factor to keep in mind is, although these groups are included with property sale transactions in the hedonic regression, they are technically a classification of people (and their typical behavioural patterns) based on factors such as ethnicity profiles, age of residents, rate of benefit claimants, population density and a great number of other factors (CACI, 2019). As such, they are not a classification of

[3] See: https://acorn.caci.co.uk/

[4] See: https://acorn.caci.co.uk/downloads/Acorn-User-guide.pdf

A Lavish Lifestyle

B Executive Wealth

C Mature Money

D City Sophisticates

E Career Climbers

F Countryside Communities

G Successful Suburbs

H Steady Neighbourhoods

I Comfortable Seniors

J Starting Out

K Student Life

L Modest Means

M Striving Families

N Poorer Pensioners

O Young Hardship

P Struggling Estates

Q Difficult Circumstances

R Not Private Household

FIGURE 6.1: CACI Acorn Classification used by the ONS Hedonic Regression [4]

properties, but a categorisation of the people expected to be living in those properties. They were included in the ONS house price index on account of being found to have some explanatory power over the value of properties in the data (ONS, 2023a).

While the CACI Acorn attribute has a match rate in excess of 99% on existing properties, the match rate for new build properties is only 40%. This is due to the fact that the Acorn dataset provided to ONS is only updated once per year, thus many new builds which are allocated newly generated postcodes have not yet been classified into an Acorn group. This may lead to increased volatility in producing the ONS house price index on new build properties (ONS, 2023a).

6.1.2 Methodology

The ONS house price index is built on a hedonic regression imputation model (see OECD, Eurostat, et al., 2013), which uses a semi-log form:

$$\log(p_i) = k + \sum_j \beta_j X_j^i + \epsilon_i$$

where:

x_j^i takes the value 1 if p_i has characteristic j, otherwise taking 0 [5]

β_j is the coefficient associated with characteristic j

p_i is the price of property i

e_i is the statistical error for p_i

k is a constant

Unlike the RPPI which uses a single twelve-month rolling time dummy hedonic regression model (see Section 3.1), the ONS house price index fits a separate regression for each month of the dataset, where property transactions in the current month are used to estimate the β_j coefficients. The *goodness-of-fit* statistic achieved is typically in the region of 80%, however, this is not always a reliable metric for the reliability of a given model, as discussed in Section 3.3.

The majority of the variables fed to the ONS house price index model are categorical dummy variables; the only exception being the floor area variable, which is a continuous value. This means that separate regressors are fit to estimate the value of having one room, two rooms, three rooms and so forth. As a result of the discrepancies between the definition of a *habitable room* discussed in Section 6.1.1.2, distinct categorical variables must be allowed for Scottish properties and homes in England & Wales, as well (ONS, 2023c).

The full list of variables included in the hedonic regression model are as follows:

[5] This is with the exception of floor area, where the value in square meters is used as the value of x_j^i

- Local Authority District (a total of 374, representing the number in the UK) [6]
- CACI Acorn Classification (a total of 18)
- Property Type (a total of 4) [7]
- Floor Area (a total of 1 continuous variable)
- Habitable Rooms [England & Wales] (a total of 8) [8]
- Habitable Rooms [Scotland] (a total of 8) [9]
- New or Existing Build (a total of 2)

Thus, we estimate that the ONS hedonic regression models contains approximately 415 regressors [10], each of which are fit on a monthly basis and acquire a fresh set of weights, according to the properties transacted in that month. The monthly house price index is then estimated by taking a fixed basket of properties, typically the entire set of properties from the year prior, and evaluating those properties on each monthly regression. This gives an estimate of what each property would be worth if sold in any given month of the time period being studied. A geometric mean is then taken of these estimates, in order to give an *average* property price for each month, which can be used to derive the index level values and the monthly percentage changes.

The mix-adjustment is somewhat implicit in the fact that, given the entire set of sales for the prior year are being used to impute prices in each time period of interest, the mix is fixed to the proportions transacted for each variable from that prior year. As stated by ONS, 2023b:

> The process of mix-adjustment requires that, in each January, a fixed basket of properties is updated to reflect changes in the composition of properties being sold. This basket is then used to produce modelled prices for

[6] See: https://geoportal.statistics.gov.uk/datasets/ons::local-authority-districts-december-2022-boundaries-uk-bfc/explore

[7] Detached, Semi-detached, Terraced and Apartment

[8] Ranging from one room up to a maximum of eight rooms (ONS, 2023b)

[9] Ranging from one room up to a maximum of eight rooms (ONS, 2023b)

[10] In practice, one variable will be omitted from each category, given that a zero in all other values implies presence of the excluded value. Thus, the total number of fittable coefficients will be 409.

> the year, before the basket is then updated again in the subsequent January. This means that the average prices produced from a fixed basket in 2016 are not directly comparable with the average price produced using the 2017 basket as they will reflect a different mix of properties.

In order to create a continuous, comparable index, the ONS calculate an adjustment factor which represents the impact of changing the mix in the month of changeover. The *average* house price output by the model is then scaled by that adjustment factor, giving a homogeneous time series.

To deal with the issue of missing or unmatched attributes, the ONS includes these records in the hedonic regression fitting process by imputing the missing value using a *nearest neighbour* approach. However, the weight of records which contain imputed values will be downweighted during the coefficient estimation, in proportion with the judged significance of the variable(s) they are missing. For example, assume that ONS have determined that the number of rooms is responsible for 20% of the explanatory power of the value of a property, then any record which is missing this value will be down-weighted to 0.8 when the regression is being fit.

6.1.3 Strengths and limitations

According to the ONS, one of the key strengths of their hedonic regression house price index is that it has wide coverage of the entire set of sale transactions, both cash and mortgage sales, through use of the real land registry dataset (ONS, 2023b). Indeed, most other privately produced house price indices in the UK utilise a limited dataset; for example, Nationwide and Halifax both produces indices based on their own mortgage approvals. This has the potential of introducing bias to the model, given the lenders may have specific selection criteria for the type of customer they typically approve, something discussed in Section 2.4.

Another strength of the model is the ability to break down to a highly granular level, for example, down to the local authority district level. This is due to ONS using the entire land registry transaction set, meaning many local authority strata have enough transactions per month to generate a somewhat reliable signal, however, this varies heavily from location to location.

One of the biggest drawbacks which the ONS House Price Index suffers from is a lengthy publication lag. According to ONS, their HPI is typically published with a lag in the region of two months, e.g. December's index is not published until the latter part of February (ONS, 2023b). Other house price indices in the UK which make use of private lender data or listed asking prices, such as Rightmove [11] offer a more timely publication, typically being able to publish within a few weeks of the end of the time period of interest.

Another key issue with this house price model lies in the handling of new build properties. Sale transactions which involve a new property necessitate the creation of a new register by the land registry, rather than a transfer of registry. As a result, they typically have a lengthier lag before appearing in the Price Paid Data or Registers of Scotland dataset. As such, ONS, 2023b found that their initial estimate for the house price index was consistently overestimating the value of new build properties for this reason. This was causing a distortion in the index, whereby the index would always be revised downward in subsequent months. ONS have determined that, despite accounting for around 10% of the total number of transactions, new builds were the driver of 80% of the monthly revisions.

In order to handle this, ONS made the decision to exclude new build transactions from the initial release of each month of the index. This would be achieved by calculating the change in the model output without new builds from month $t-1$ to month t and applying that change to the model output *with* new builds for month $t-1$. New builds are then returned to the month t model value in each subsequent month. The results of this change are shown in Figure 6.2. While the updated methodology produces a somewhat closer result to the final estimate, it is still rather volatile.

This issue with new builds has become a larger concern since the index was resumed following the *COVID-19* pandemic, as the processing of new build transactions has slowed down. As a result, the ONS have needed to pool new build transactions across multiple months, effectively smoothing the new build component of the house price index, as the Central Statistics Office did with the entire RPPI (discussed

[11] See: https://www.rightmove.co.uk/news/house-price-index/

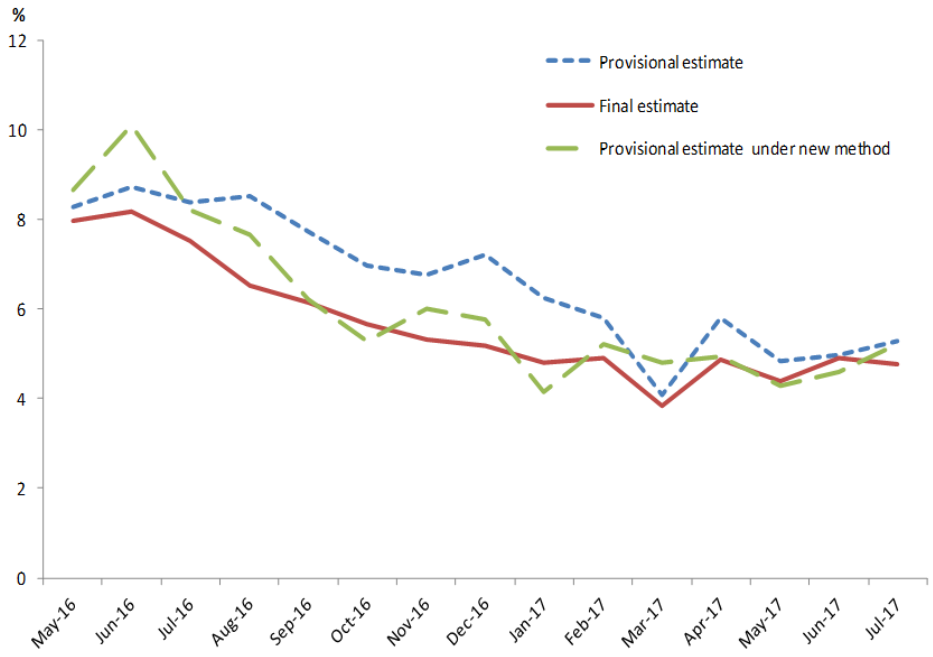


FIGURE 6.2: Comparison of ONS index with new build methodology change

in Section 3.1.1). The pooling mechanism as described in the most recent house price index report as of the time of writing is as follows [12]:

- June 2022 includes new build transactions from May 2022 and June 2022
- July 2022 includes new build transactions from June 2022 and July 2022
- August 2022 includes new build transactions from July 2022 and August 2022
- September 2022 includes new build transactions from August 2022 and September 2022
- October 2022 includes new build transactions from September 2022 and October 2022
- November 2022 includes new build transactions from October and November 2022

[12] See: https://www.gov.uk/government/collections/uk-house-price-index-reports

6.2 Applying our stratified model to the UK Price Paid Dataset

In this section, we will explore the Price Paid Data in more depth, as this is the dataset which we will be using in order to test our stratified, mix-adjusted median model on the UK. Furthermore, we will outline a tweak to our methodology which is intended to deliver additional smoothness over our previous incarnation of the stratified index algorithm.

6.2.1 Dataset

6.2.1.1 Specification

As mentioned previously, the Price Paid Dataset is made publicly available on a monthly basis by HM Land Registry. In addition to our geospatial stratification model, we also have the property type at our disposal, which can be used as an additional variable to encode in the geohash$^+$, allowing more granular stratification. We do not have access to any additional, explicit characteristics of the property such as the number of bedrooms, floor area or neighbourhood quality, as the ONS hedonic regression model does.

Unfortunately, we also do not have access to the paid Registers of Scotland data for this analysis. However, we expect this omission to have only a minor impact on the ability of our model to achieve robust performance, given that Scotland only accounts for around 13.5% of the total number of sales in the UK per annum (as of the time of writing) [13].

6.2.1.2 Characteristics: Transaction Volume

As with our prior analyses, before assessing the performance of our house price index algorithm, it is useful to survey some of the basic characteristics of the Price Paid Data, in order to get a baseline view on the dataset. In terms of transaction volume, the Price Paid Dataset offers a much richer monthly sample for our model to exploit than the datasets we have looked at previously. Figure 6.3 shows the monthly transaction volume over a period in excess of ten years, with the stacks on

[13] See: https://www.gov.uk/government/statistics/uk-house-price-index-for-december-2022/uk-house-price-index-summary-december-2022

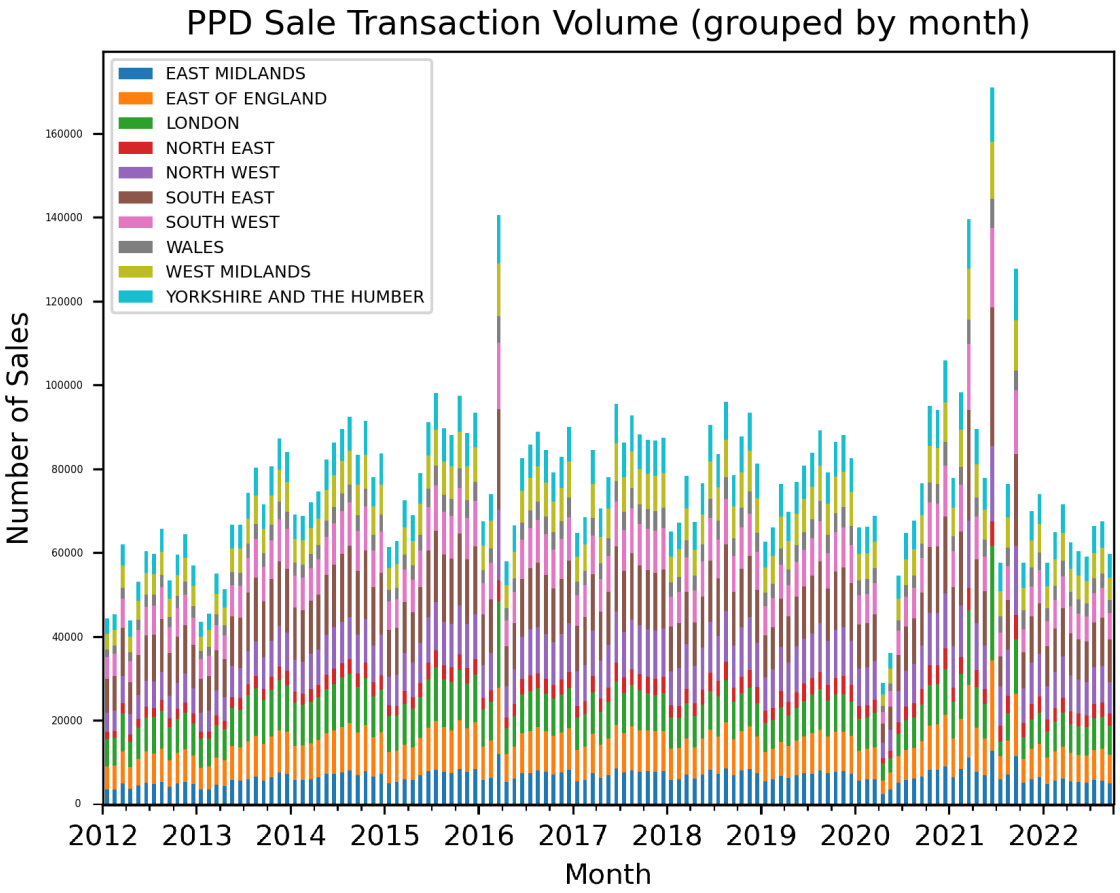


FIGURE 6.3: Price Paid Data: Volume from 01-2012 to 09-2022 (inclusive), broken down by region

each monthly bar illustrating the regional breakdown within each month, for Wales and the nine regions of England. The monthly sales volume ranges from a low of around 30,000 in April 2020, to a high of around 175,000 transactions in June 2021.

Table 6.1 shows that even the region with the lowest mean number of transactions per month, the North East, has approximately 3,460 sales. This is substantially higher than the total average monthly transaction count for the entirety of Ireland, meaning that our model should have ample sale record volume to produce a signal for every sub-region of England & Wales.

Furthermore, seasonal analysis on the transaction volumes in Figure 6.4 reveals that a seasonal pattern in transaction volume was detected (with stable residuals) for the majority of our sample period, however, this has broken down since the onset of the *COVID-19* pandemic in March 2020. Prior to that, it was clear that home

TABLE 6.1: Price Paid Data: Transaction Volumes per region

Region	Mean	Median	St.Dev
North East	3,459	3,513	841
Wales	3,975	4,050	981
East Midlands	6,812	6,693	1,664
West Midlands	7,213	7,081	1,813
Yorkshire And The Humber	7,271	7,219	1,767
South West	8,806	8,789	2,314
East Of England	9,015	9,004	2,351
London	9,396	9,293	2,717
North West	9,825	9,818	2,545
South East	13,355	13,124	3,557

buying was most popular from Easter onward, through summer; experiencing a sharp drop each January and March. It appears that this fall in January transactions is due to a slowdown in sales approaching the holiday season in December, owing to the typical one-month lag to a sale settling (Norman Mille and Pampulov, 2013). A possible explanation for the spike in transactions each February then, is the clearing of backlogged sales which did not complete prior to people taking holidays, most of which would then be settled by HM Land Registry during February.

We can also analyse the transaction volumes separated by property type. Figure 6.5 demonstrates the monthly sale volume, with the bar stacks representing each of the four property types in the Price Paid Data; detached, semi-detached, terraced and apartment. Again, we can see that the mix of properties is such that our model

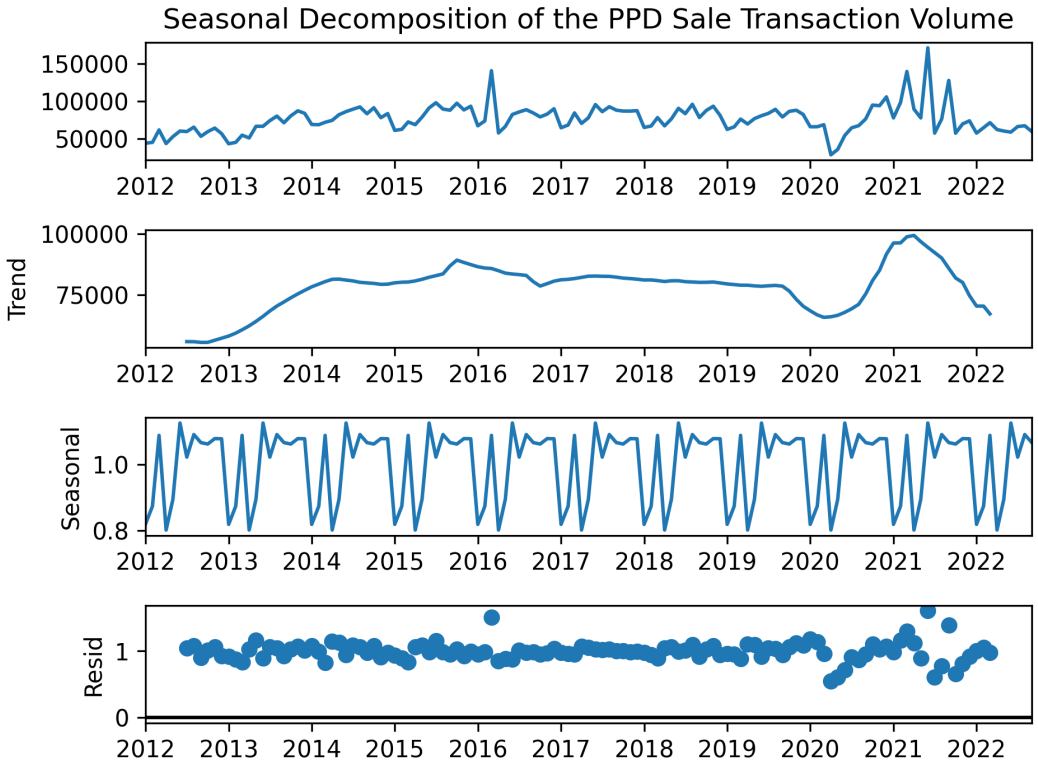


FIGURE 6.4: Price Paid Data: Volume Seasonality from 01-2012 to 09-2022 (inclusive)

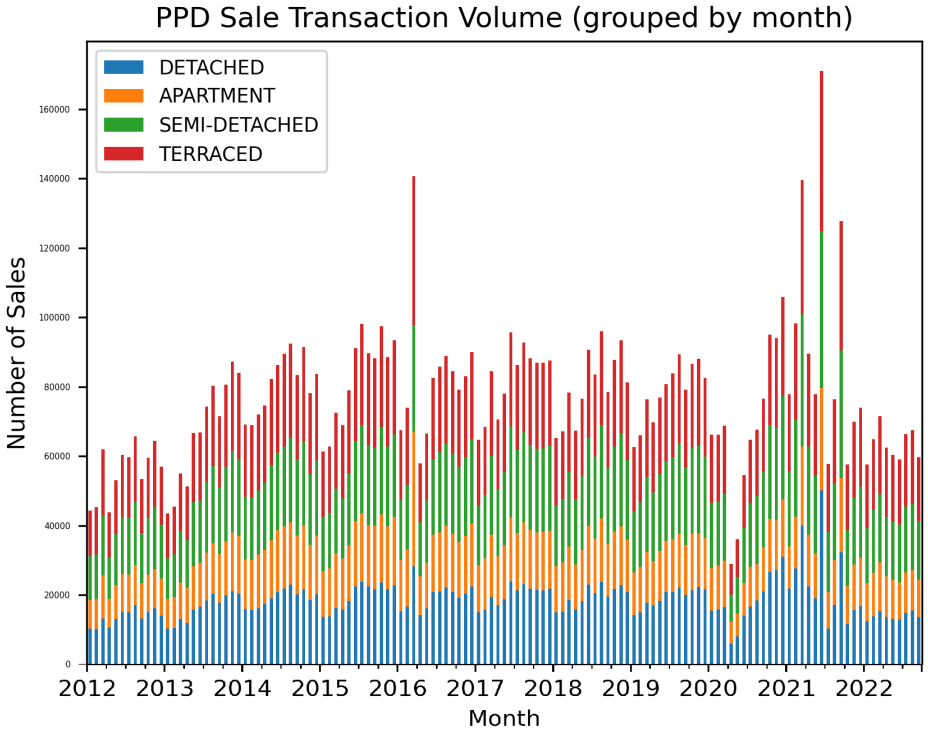


FIGURE 6.5: Price Paid Data: Volume from 01-2012 to 09-2022 (inclusive), broken down by property type

TABLE 6.2: Price Paid Data: Transaction Volumes per property type

Property Type	Mean	Median	St.Dev
Apartment	14,519	14,481	3,910
Detached	18,444	18,232	5,372
Semi-Detached	20,760	21,238	4,901
Terraced	22,307	22,244	5,092

will have an ample transaction set to calculate sub-indices for each property type. Table 6.2 shows that the typical volume in the smallest category, apartments, is over six times the total monthly transaction pool from our original sparse and frugal model (see Chapter 3).

The final potential attribute for stratification in the Price Paid Data is the build type: new build or existing. The share of new build properties is quite small, relative to the number of existing properties; taking up around 10% of the total pool, as seen in Table 6.3. Figure 6.6 demonstrates the monthly sale records with the bar stacks representing the build type. Despite composing a comparatively low share of the total number of transactions, it is still considerably in excess of the data volume used in our original model.

However, evidence from recent months supports the difficulty which the ONS have had in handling new build regression estimation; necessitating their methodological change of pooling of multiple months of sales (as discussed in Section 6.1.3). Indeed, Table 6.4 demonstrates that new build volumes since April 2022 have dropped by 93.3%, from a mean of 7,581 transactions per month over our entire sample period, to a mean of 507 per month from April 2022 to September 2022 (inclusive). This number has deteriorated further in even more recent months, with an average of 273 sales per month in the last three months of the sample. We will explore the impact of this in greater depth when analysing our results.

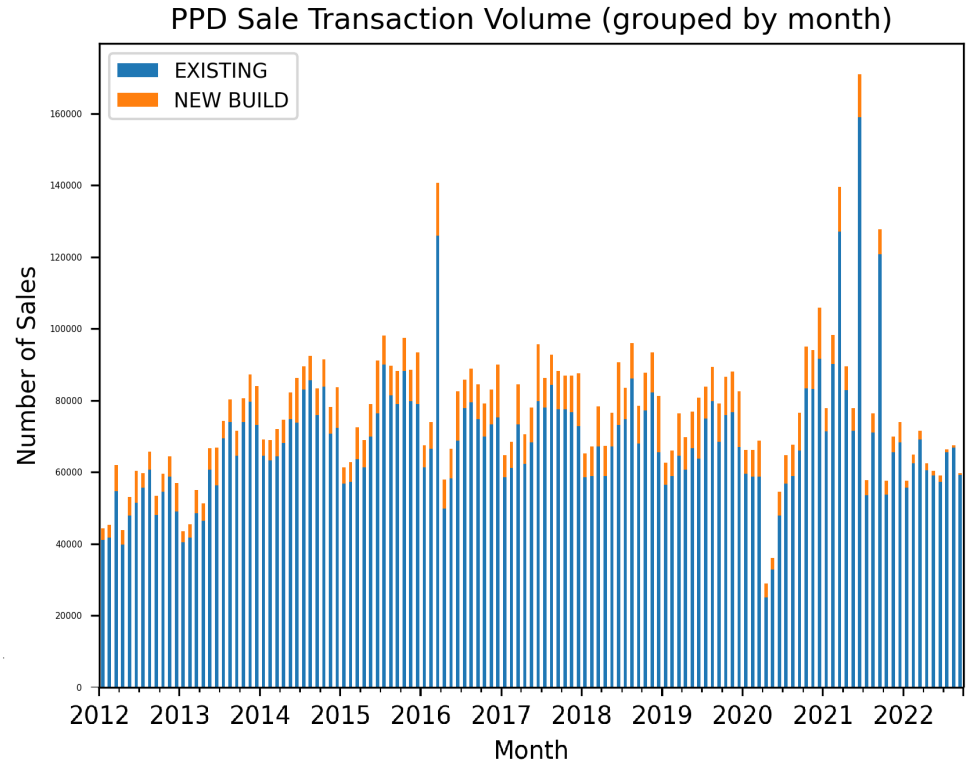


FIGURE 6.6: Price Paid Data: Volume from 01-2012 to 09-2022 (inclusive), broken down by build type

TABLE 6.3: Price Paid Data: Transaction Volumes per build type

Build Type	Mean	Median	St.Dev
New Build	7,581	7,599	3,719
Existing	68,449	67,160	17,042

TABLE 6.4: Price Paid Data: Transaction Volumes per build type (since April 2022)

Build Type	Mean	Median	St.Dev
New Build	507	478	276
Existing	61,262	59,701	3,897

6.2.1.3 Characteristics: Sale Price Distribution

In order to set some baseline expectations for the volatility of the data, we can explore the sale price distribution of the Price Paid Data across a number of strata. First, looking at a regional breakdown, Figure 6.7 shows a considerable difference in the median price level across different parts of the UK. Thus, if a mix-adjusted national median was taken without stratification, much of the behaviour of the price distribution of more expensive regions such as London and cheaper regions such as Wales would have suppressed influence on the overall price index, as discussed in Section 3.2.2.3.

Figure 6.8 demonstrates the difference in distribution between the four distinct property types. Interestingly, the distribution would suggest that apartments have been growing in value more slowly than other types of properties, which coincides with London's growth appearing to stall in Figure 6.7, given that London has a high share of apartments [14]. The chart also shows that the interquartile range for each kind of property is rather broad, with every property type showing a range in excess of £200,000 in recent times. As a result, we would expect significant volatility pass-through to basic mean and median price indices.

[14] See: Appendix A - Figure A.3

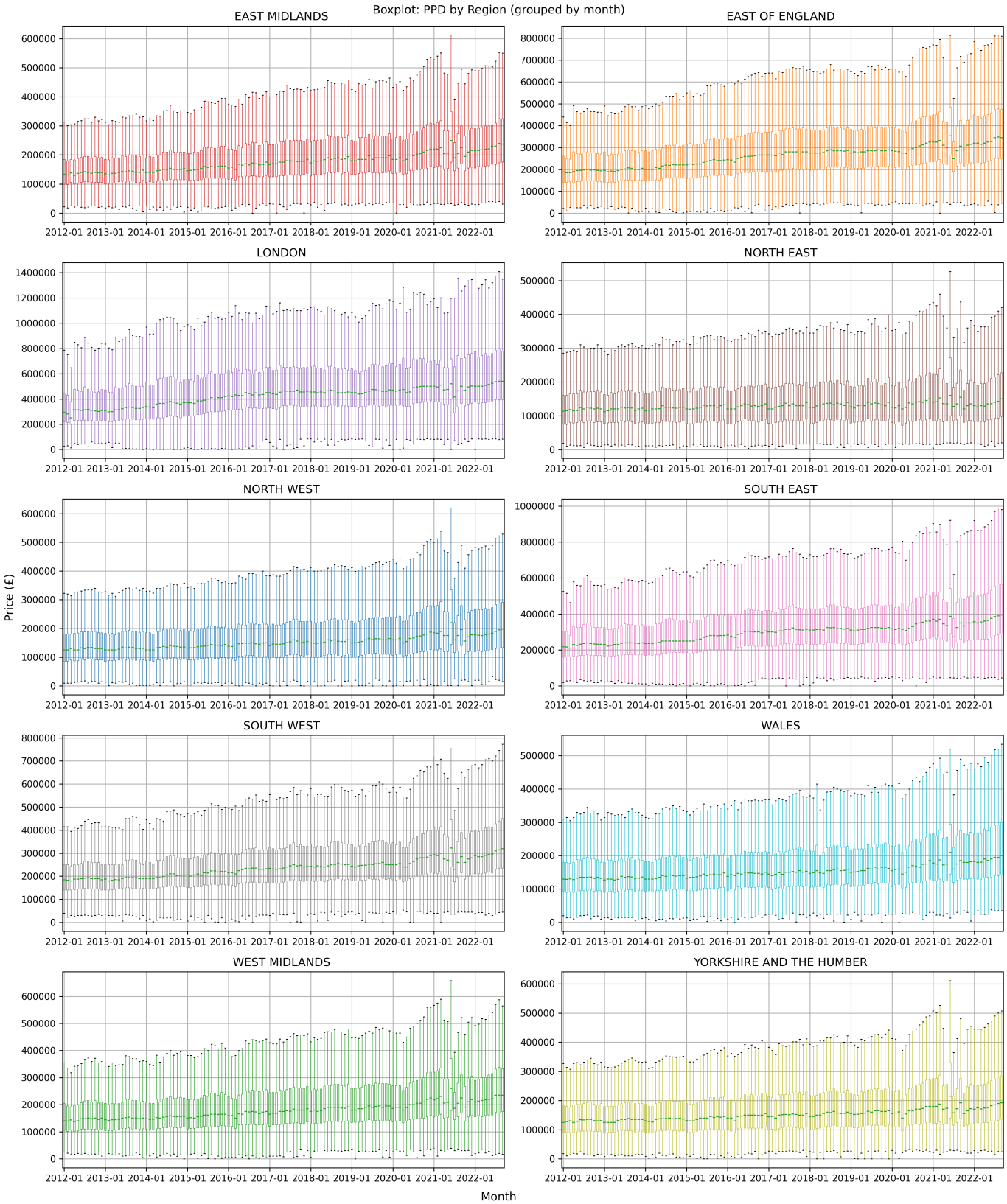


FIGURE 6.7: Price Paid Data: Price Distribution from 01-2012 to 09-2022 (inclusive), broken down by region

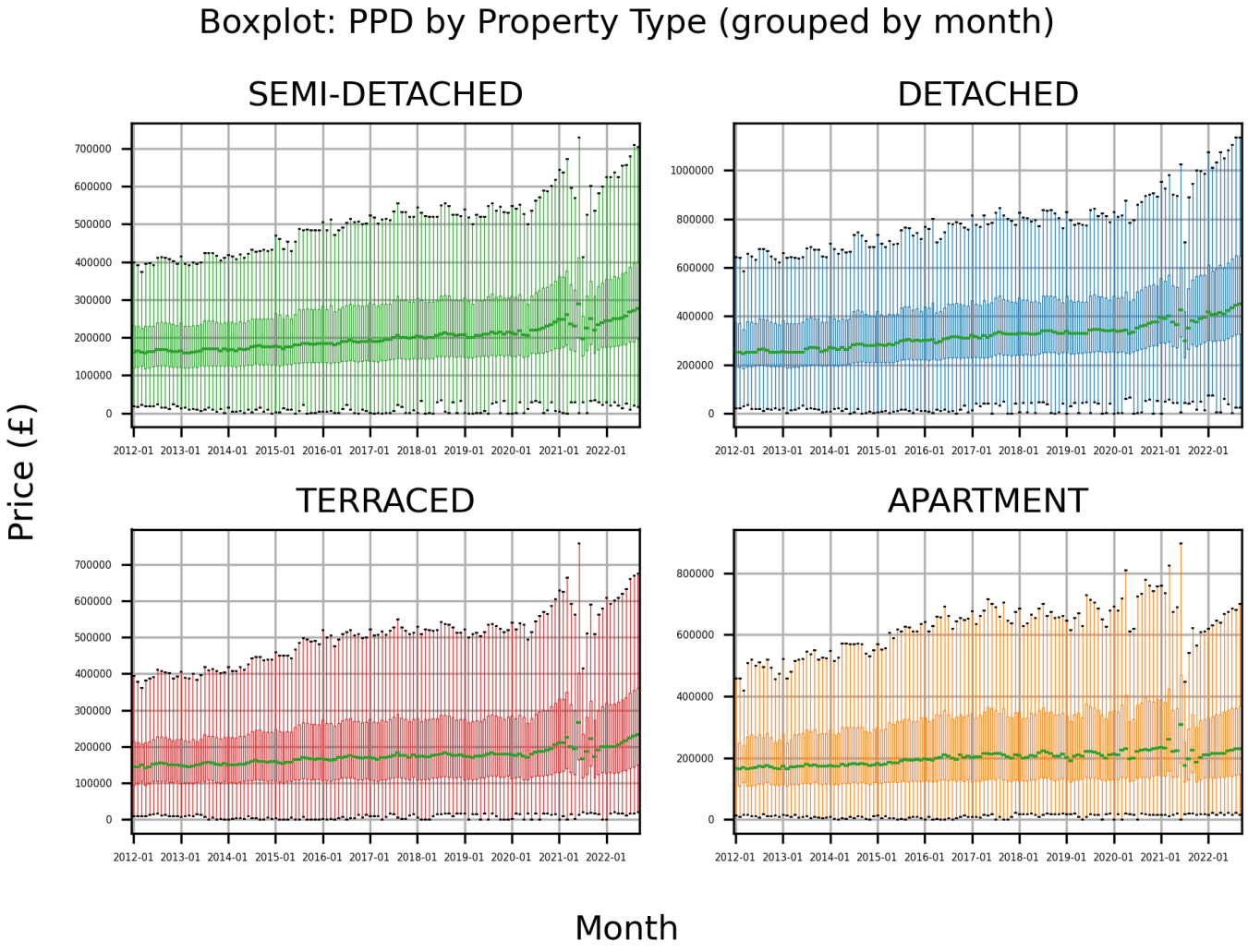


FIGURE 6.8: Price Paid Data: Price Distribution from 01-2012 to 09-2022 (inclusive), broken down by property type

Additional transaction price distribution charts are included in Appendix A, demonstrating the dispersion of prices across the dataset as a whole and broken down by build type.

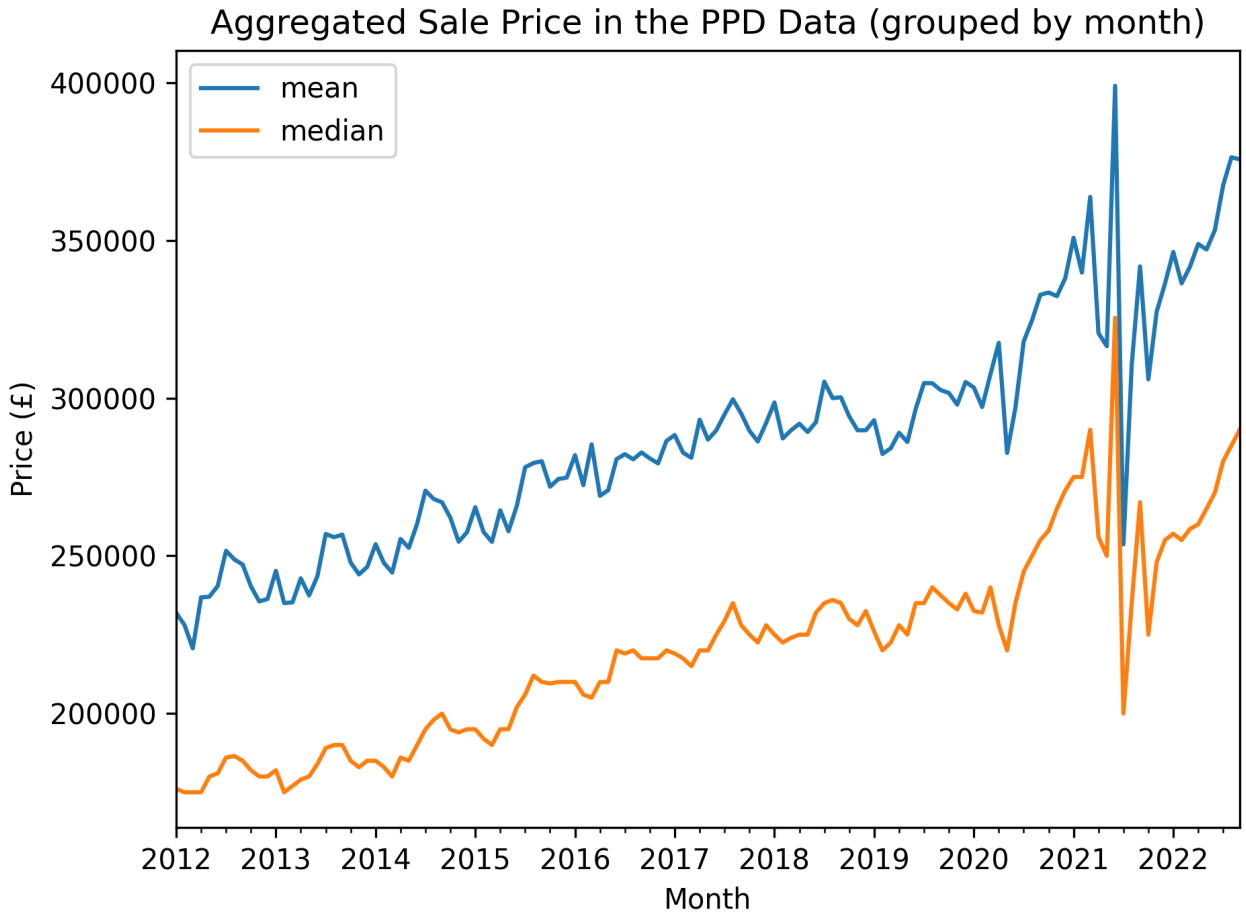


FIGURE 6.9: Price Paid Data: Mean/Median Price from 01-2012 to 09-2022 (inclusive)

6.2.1.4 Mean and Median Price Indices

We can generate naive, un-adjusted mean and median price indices across the various strata, in order to assess the level of noise captured by these simple measures. One would expect that the noise should be somewhat lower on this dataset when compared with the Property Price Register data (explored in Section 3.2.1), due to the larger number of samples in a typical monthly sample.

Despite this, Figure 6.9 still shows an unrealistic amount of volatility in both the mean and median national indices, versus what would be expected from natural property market behaviour. The mean price index shows an average absolute monthly change of 3.05%, with a standard deviation of 4.60% (on the absolute values), while the median index averages 2.56%, with 4.90% standard deviation. We will explore our previously proposed smoothness metrics (see: Section 5.2) further

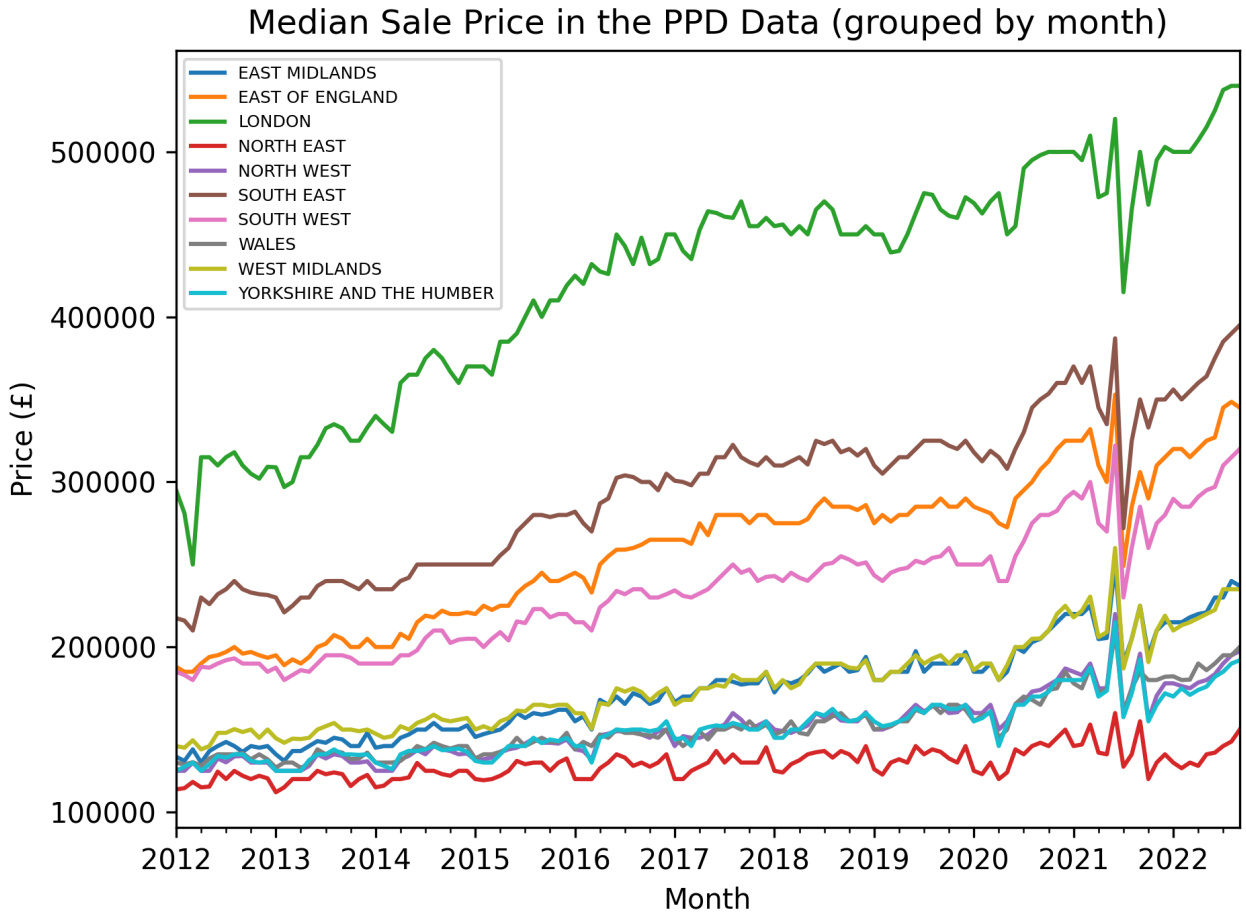


FIGURE 6.10: Price Paid Data: Mean/Median Price from 01-2012 to 09-2022 (inclusive), broken down by region

when analysing index results later in this chapter. The year of 2021 was particularly volatile, showing a surge of 26% in June, followed by a drop of 36.4% in July, rounded out with a jump of 22.5% in August. According to Agnello and Schuknecht, 2011; Leung and Tsang, 2013, typical housing market behaviour consists of long-run cycles where prices move smoothly with momentum. Thus, price action of these extreme magnitudes is clearly not indicative of housing market dynamics and is likely due to noise or sample bias in the transactions reported during the periods in question.

Figure 6.10 demonstrates the median price index on a region-by-region basis. As expected, this stratification of the sample aids somewhat in reducing noise on the naive, un-adjusted median index, however, the volatility present is still significant, particularly in certain regions.

Interestingly, the wild behaviour of the index during 2021 observed previously

appears to be more subdued in the regions with a lower median value, even in relative terms. Table 6.5 shows the average absolute monthly percentage change on both the mean and median index, broken down by region, alongside the standard deviation of the monthly changes.

TABLE 6.5: Mean and Median Index: Average Absolute Monthly Change, by region

Region	Mean Index[a]	St.Dev[b]	Median Index[a]	St.Dev[b]
EAST MIDLANDS	2.53	137.67	2.9	123.17
EAST OF ENGLAND	2.2	154.75	2.07	165.48
LONDON	3.51	89.55	2.38	145.8
NORTH EAST	3.55	123.94	3.82	102.08
NORTH WEST	2.9	168.51	2.85	143.95
SOUTH EAST	2.42	128.87	2.13	166.66
SOUTH WEST	2.68	131.94	2.24	155.13
WALES	2.49	109.64	2.79	114.63
WEST MIDLANDS	2.76	133.92	2.82	138.42
YORKSHIRE AND THE HUMBER	2.88	146.96	2.92	141.79

[a] Values quoted are the mean of the absolute monthly percentage changes of each index

[b] Standard deviation is reported as a percentage of the quoted average index values

Section A.3 includes some additional mean and median price indices, with stratification by property type and by build type.

6.2.2 Methodological alterations and improvements

6.2.2.1 Adjustment of neighbour weights

As discussed in Section 3.2.2.3, the primary issue with any approach of modelling property prices which solely focuses around the median, is ignorance towards the behaviour of the distribution above and below said median. As such, properties with values which sit further away from the median have little influence on the median level, while those near it have a disproportionate impact.

In order to solve this, we introduced a new model whereby each property receives a distinct stratification base, allowing each individual property to have an equal contribution to the aggregate index. In Section 4.6.1, we introduced the GeoTree, which allowed us to speed up our nearest neighbour searches dramatically, by instead querying a bucket of approximate neighbours of each property. As a result of this, we had a number of potential stratification bases per property, rather than a single base.

Previously, we took the median of the neighbours to be the stratification base and used the price change ratio generated by that property as the value passed to the aggregation stage of the index computation. However, there is a methodological improvement which can be made, whereby we use the additional information available to reduce the noise in the model further. Referring back to our prior discussion, we now have a scenario whereby we are taking the median neighbour as a stratification base, while ignoring all of the information about the distribution of neighbours of said property. This is an identical argument to our basis for introducing localised stratification on the monthly transaction level, just on a deeper level in the algorithm.

Formally, our original methodology was such that the change ratio for each property was calculated as:

$$R(p_i) = \left(\frac{price(p_i)}{MEDIAN(\{price(p_n) \mid p_n \in N(p_i)\})} \right) - 1$$

where $R(p_i)$ is the price change ratio for p_i and $N(p_i)$ is the set of approximate

neighbours of p_i, given by the GeoTree. However, we can include each of the neighbours in the model aggregation stage by returning the change ratio of every neighbour of p_i, scaled by a corresponding weight factor according to the total number of neighbours.

As such, our weighted change ratios for p_i are defined as:

$$R(p_i) = \left\{ \frac{1}{|N(p_i)|} \cdot \left(\frac{price(p_i)}{price(p_n)} - 1 \right) \middle| \, p_n \in N(p_i) \right\}$$

The price ratios for p_i will be aggregated with a set of price ratios from every other property in the monthly sample before the weighted average of all price ratios will be computed. This methodological change allows us to account for the price action of every neighbour around each property returned by the GeoTree, rather than solely considering the median neighbour; applying our concept of accounting for the behaviour of the entire price distribution a layer deeper in the model algorithm, in order to reduce noise further.

6.3 Stratifying our *GeoPrice* index using geospatial data

Initially, we will perform the stratification process of our *GeoPrice* model only using geospatial matching on the geohashes associated with each property. Each property in the dataset has been matched with GPS co-ordinates using the public postcode mapping dataset, maintained by the Office for National Statistics [15]. Once matched with a pair of GPS co-ordinates, each property has then been allocated a corresponding geohash, as described in Section 4.3.1.

Further to fitting our model on the national level, we will assess our ability to generate regional sub-indices for each of the previously mentioned regions of England, plus Wales.

[15] See: https://www.ons.gov.uk/methodology/geography/geographicalproducts/postcodeproducts

6.3.1 National index

6.3.1.1 Time series analysis

Figure 6.11 shows our *GeoPrice* house price index with the ONS hedonic regression model superimposed on the chart. As evident from the chart, both of the indices follow a highly similar trend over our sample period of over ten years. However, the ONS house price index experiences significantly more month-to-month volatility throughout the entire data history. We will quantify this more formally when assessing our smoothness metrics in Section 6.3.1.2.

The volatility differential is better illustrated through Figure 6.12, which shows the monthly percentage changes of each index superimposed. In terms of index similarity, our model shows a Pearson's r correlation coefficient of **0.8588** with the ONS hedonic regression on the monthly change values, corroborating the findings of McDonald, Smith, et al., 2009; Prasad and Richards, 2008; sufficiently stratified models can achieve high levels of congruence with hedonic regression models.

As discussed in Section 6.2.1.3, the middle of 2021 exhibited some highly volatile behaviour in the naive un-adjusted mean and median price indices. In the ONS House Price Index, we can see that this noise was suppressed, with June 2021 surging 5.73%, July 2021 dropping 4.79% and August 2021 increasing 2.95%, for a compounded change of 3.63% over the three month period. Our *GeoPrice* house price index improved significantly on this noisy signal, posting +1.91%, -0.59% and +1.23% on the three respective months, for a compounded change of 2.56%. As evident in Figure 6.11, the ONS index overshoot of this cumulative increase came back in line with the *GeoPrice* index in the following three months, however, the ONS index has increasingly diverged beyond that time. We will explore the reasoning behind this disparity in greater depth in Section 6.3.2.

6.3.1.2 Smoothness metrics

In Table 6.6, we demonstrate a variety of smoothness metrics for the naive mean, naive median, ONS and *GeoPrice* house price indices. By every metric, the *GeoPrice* house price index exhibits a significantly smoother index profile than the ONS house price index, while achieving a high degree of correlation and reporting a very similar

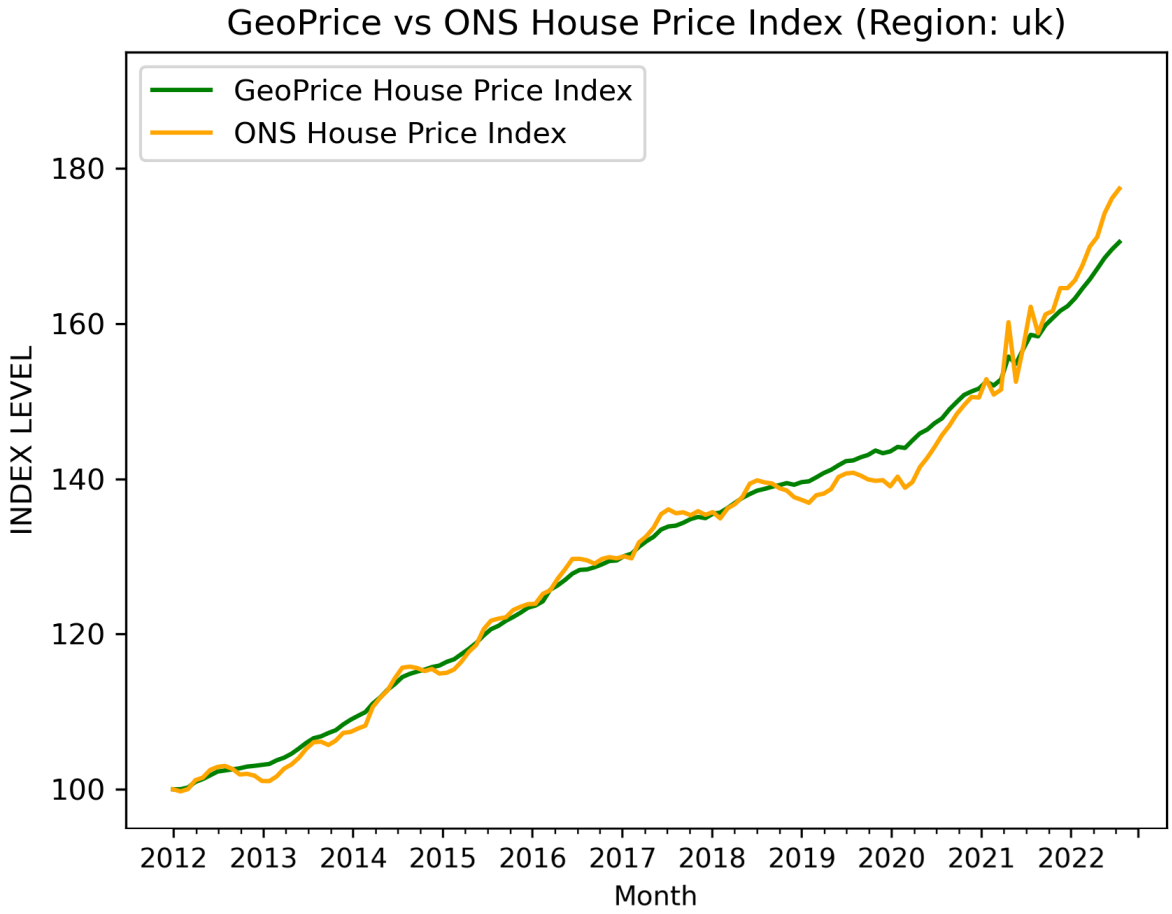


FIGURE 6.11: ONS vs *GeoPrice* House Price Index [UK] from 01-2012 to 09-2022

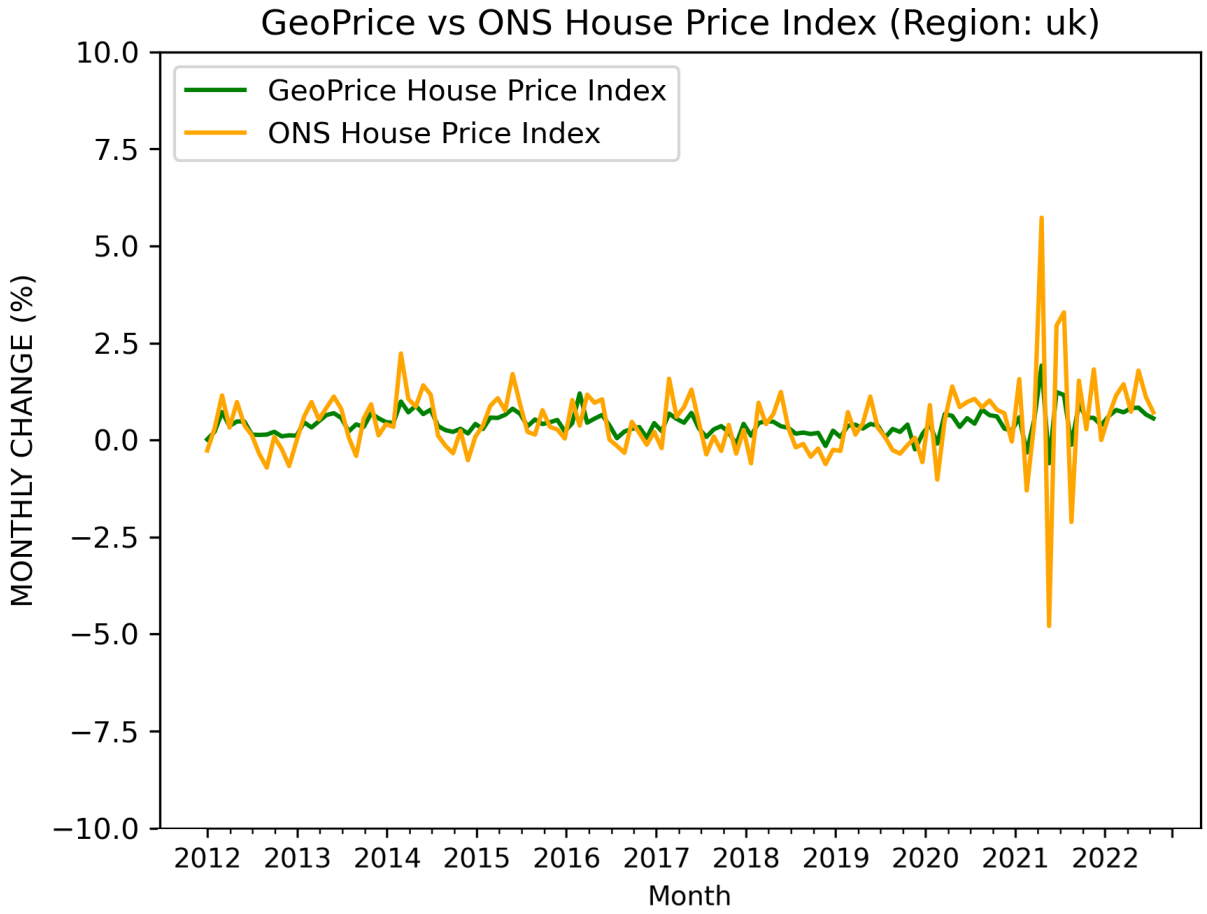


FIGURE 6.12: ONS vs *GeoPrice* House Price Monthly Change (%) [UK] from 01-2012 to 09-2022

level of total price appreciation across the sample period of over ten years. The standard deviation of the differences is 72.15% lower in the *GeoPrice* Index versus the ONS, while the mean spike magnitude is 92.7% smaller.

The naive mean and median indices show highly volatile performance and only serve to add contextual background to the *GeoPrice* and ONS house price index results.

TABLE 6.6: Smoothness Metrics for Mean, Median, ONS and *GeoPrice* Indices [UK]

Model	Mean[a]	Median[a]	St. Dev[b]	Min[b]	Max[b]	St. Dev of Diffs	MSM
GeoPrice [c]	0.44	0.41	0.32	-0.6	1.92	**0.44**	**0.82**
ONS HPI	0.75	0.54	1.0	-4.79	5.73	1.58	11.24
MEAN	3.05	1.99	5.5	-36.44	26.09	9.63	444.95
MEDIAN	2.57	1.32	5.51	-38.56	30.2	9.66	462.45

[a] Values quoted are the mean of the absolute monthly percentage changes of each index
[b] Values are reported in percent
[c] *GeoPrice* algorithm is run with geospatial stratification, only

6.3.2 Issues with New Build transactions

A number of issues with new build transactions has caused the ONS to alter methodology in an attempt to work around the matter, as discussed in Section 6.1.3 and our analysis of transaction volumes in Section 6.2.1.2. The pooling of new build transactions across multiple months in recent times is a significant issue. Firstly, if the small sample of new build properties in each month happen to show a strong movement in either direction, this will be broadcast across multiple months through this pooling method. Furthermore, it is unclear whether ONS have made adjustments to the fitting or weighting process, in order to account for the fact that these new build transactions are being repeated.

It is curious then that the recent, larger divergence in the *GeoPrice* and ONS indices observed in Section 6.3.1 began to occur around the same time period in which the pooling of these new build transactions commenced. In order to judge whether

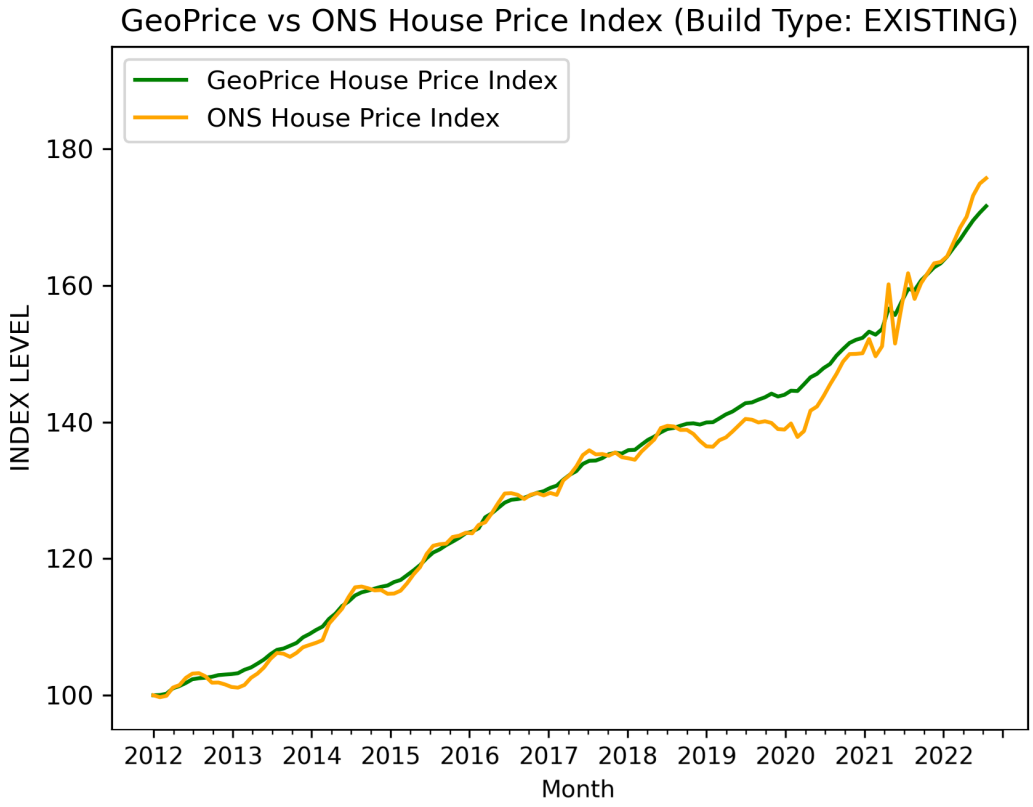


FIGURE 6.13: ONS vs *GeoPrice* House Price Index (excluding new builds) [UK] from 01-2012 to 09-2022

this could be attributed wholly or partly to this pooling behaviour, we ran our model on the subset of national transactions which consist only of existing property sales. Thus, new builds are excluded. In order to set a comparable benchmark, the ONS house price index used in this comparison is also the index which excludes new builds.

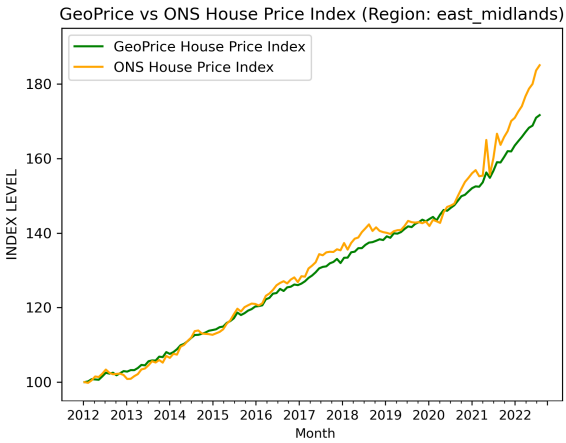
TABLE 6.7: Smoothness Metrics for ONS and *GeoPrice* Indices (excluding new builds) [UK]

Model	Mean[a]	Median[a]	St. Dev[b]	Min[b]	Max[b]	St. Dev of Diffs	MSM
GeoPrice [c]	0.45	0.41	0.33	-0.59	1.96	0.45	0.87
ONS HPI	0.77	0.59	1.08	-5.41	6.0	1.68	13.61

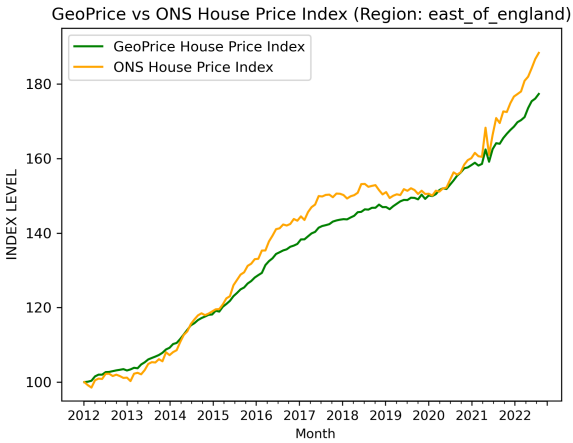
[a] Values quoted are the mean of the absolute monthly percentage changes of each index
[b] Values are reported in percent
[c] *GeoPrice* algorithm is run with geospatial stratification, only

As Figure 6.13 demonstrates, the house price indices consisting of only existing properties have reduced divergence during the period in question. Furthermore, this variant of the *GeoPrice* index is highly similar to the previously presented version, which included new builds. Thus, it seems likely that the recent ONS methodological changes related to pooling new builds are one of the drivers behind the recently observed increased divergence in the models. In the interest of completeness, Table 6.7 shows the smoothness metrics of each of the restricted models.

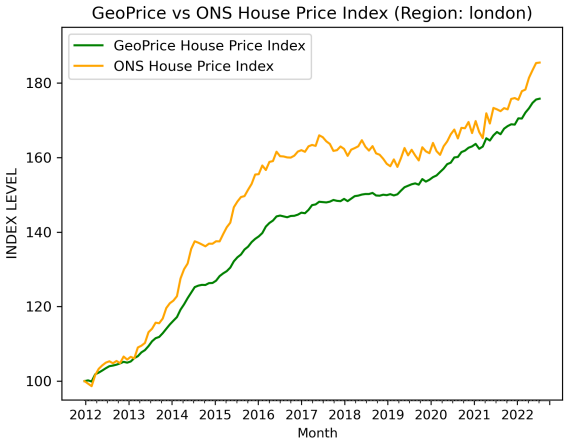
6.3.3 Regional sub-indices



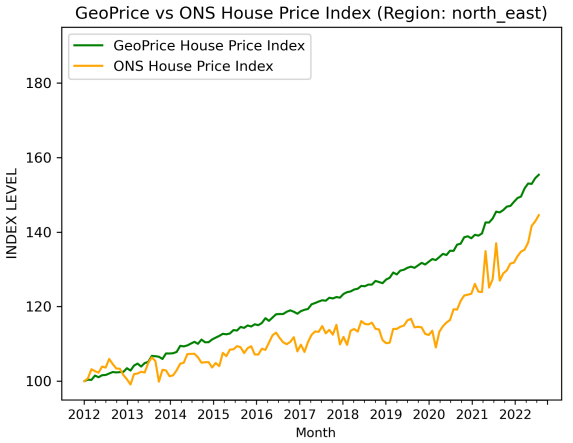
(A) East Midlands Index



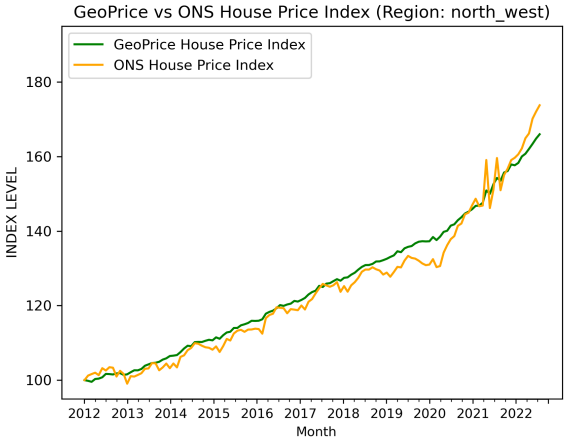
(B) East Of England Index



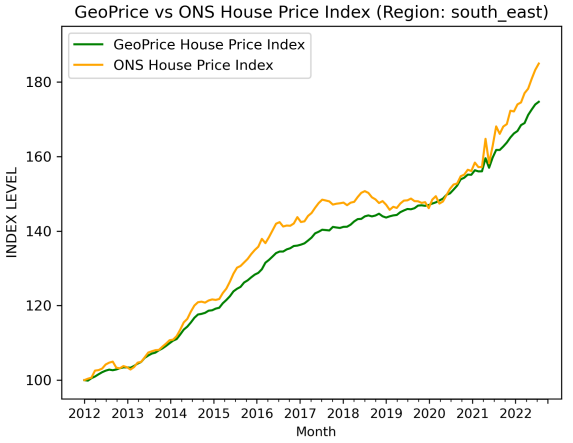
(C) London Index



(D) North East Index

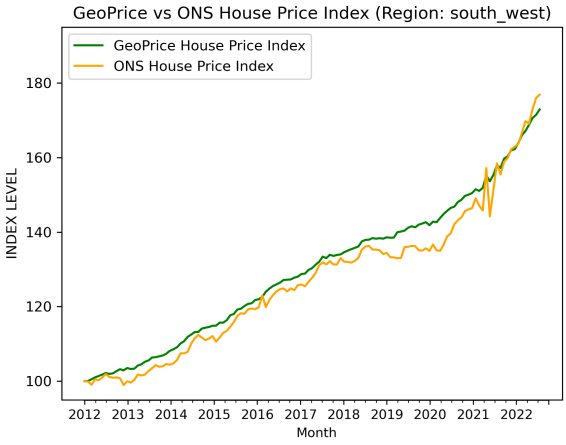


(E) North West Index

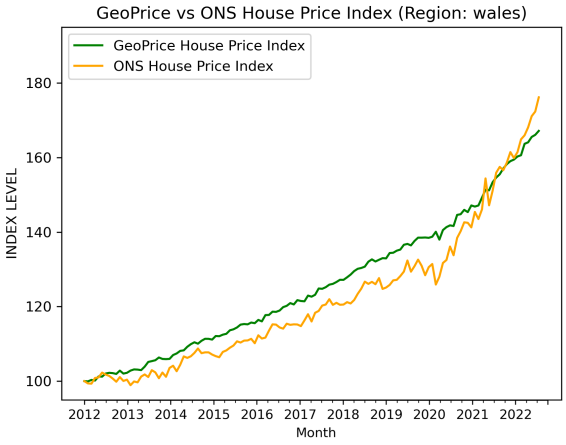


(F) South East Index

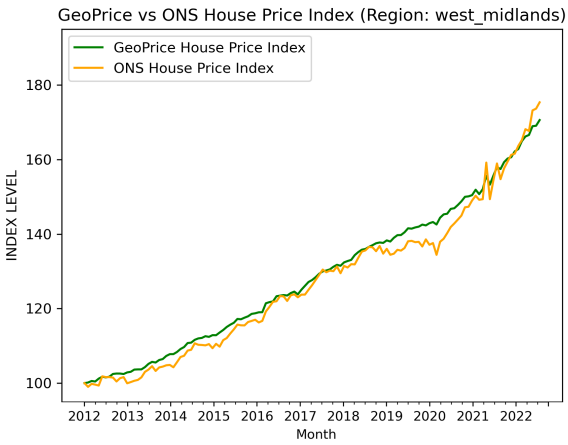
FIGURE 6.14: ONS vs *GeoPrice* House Price Index from 01-2012 to 09-2022, per region



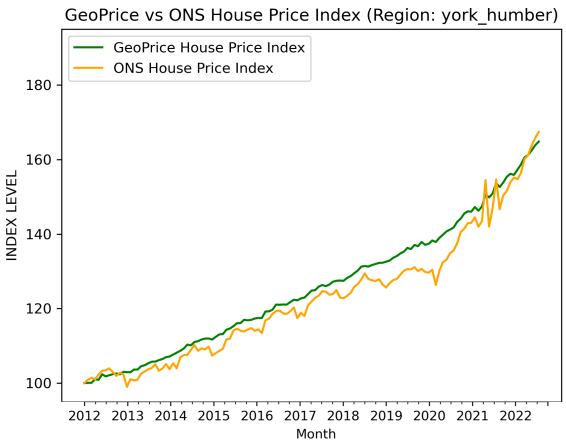
(G) South West Index



(H) Wales Index



(I) West Midlands Index



(J) Yorkshire and the Humber Index

FIGURE 6.14: ONS vs *GeoPrice* House Price Index from 01-2012 to 09-2022, per region (continued)

6.3.3.1 Time series analysis

Figure 6.14 demonstrates the sub-index generated by the *GeoPrice* model for each region of England & Wales, superimposed with the corresponding ONS hedonic regression index for the respective region. Again, we see a pattern of similar trend across all of the regions, perhaps with the exception of the North East. Indeed, Table 6.8 demonstrates a high degree of correlation in the monthly change values of most of the sub-indices. Interestingly, the two lone sub-indices with a poor level of correlation happen to be Wales and the North East, which are also the two regions with the lowest count of typical monthly transactions, as per Table 6.1. The next

lowest region, East Midlands, has over 70% more transactions than these sparsely transacted areas.

Further illustrations of the sub-index performance, including monthly change charts, can be found in Section A.4.1.1.

TABLE 6.8: Correlation of *GeoPrice* to ONS HPI, per region

Region	Pearson's *r*:
East Midlands	0.65
East Of England	0.75
London	0.67
North East	0.4
North West	0.69
South East	0.74
South West	0.65
Wales	0.24
West Midlands	0.73
Yorkshire and the Humber	0.66

6.3.3.2 Smoothness metrics

Corroborating the smoothness illustrated in the sub-index charts, Table 6.9 indicates the smoothness statistics of each sub-index, versus the ONS HPI equivalent. The mean spike magnitude is at least 75% lower in each region, with the majority of the regions seeing a reduction of over 85%. On the other hand, the standard deviation of the differences is at least 50% improved on every region in the sample, with most dropping by over 65%.

As discussed in the previous section, the lone two regions which were poorly correlated with the ONS house price index were the two regions with the lowest number of typical transactions; Wales and the North East. While the Wales index retains a similar trend over a long period, with the deviation coming from short-term monthly changes; the North East index looks significantly different. As the smoothness metrics show, this region is by far the most volatile of the ONS regional indices, with the standard deviation of the differences being over 25% worse than the next-worst region, while the mean spike magnitude is over 62% poorer. It thus appears that this volatility in the ONS index for North East is the driver of the low correlation in that region, potentially due to a poor regression fit on the lower monthly samples.

TABLE 6.9: Smoothness Metrics for ONS and *GeoPrice* Indices, in each region [UK]

Model	Mean[a]	Median[a]	St. Dev[b]	Min[b]	Max[b]	St. Dev of Diffs	MSM
East Midlands / *GeoPrice*	0.53	0.49	0.48	-0.92	1.73	**0.79**	**2.62**
East Midlands / ONS HPI	0.85	0.7	1.12	-5.67	6.16	1.84	14.11
East Of England / *GeoPrice*	0.54	0.46	0.5	-2.03	2.44	**0.78**	**2.62**
East Of England / ONS HPI	0.82	0.59	1.02	-4.32	4.87	1.61	11.25
London / *GeoPrice*	0.52	0.45	0.47	-0.8	1.93	**0.58**	**1.44**
London / ONS HPI	0.99	0.83	1.16	-1.75	4.03	1.72	12.55
North East / *GeoPrice*	0.49	0.41	0.53	-0.75	2.12	**0.87**	**3.14**
North East / ONS HPI	1.45	1.17	2.1	-7.32	8.87	3.48	53.76
North West / *GeoPrice*	0.47	0.39	0.44	-0.6	2.25	**0.71**	**2.25**
North West / ONS HPI	1.12	0.88	1.65	-8.1	8.32	2.77	32.92
South East / *GeoPrice*	0.49	0.43	0.44	-1.6	2.25	**0.63**	**1.92**
South East / ONS HPI	0.8	0.66	0.97	-3.99	4.79	1.48	9.26
South West / *GeoPrice*	0.5	0.41	0.45	-1.0	2.18	**0.69**	**1.99**
South West / ONS HPI	0.96	0.65	1.46	-8.24	7.8	2.46	28.79
Wales / *GeoPrice*	0.52	0.39	0.54	-1.51	2.1	**0.89**	**3.56**
Wales / ONS HPI	1.15	1.01	1.41	-4.67	5.69	2.26	20.7
West Midlands / *GeoPrice*	0.5	0.42	0.5	-1.52	2.35	**0.8**	**2.69**
West Midlands / ONS HPI	0.94	0.8	1.27	-6.16	6.57	2.15	18.25
Yorkshire H[c]/ *GeoPrice*	0.48	0.4	0.47	-0.68	2.25	**0.76**	**2.48**
Yorkshire H[c]/ ONS HPI	1.14	0.81	1.64	-8.07	7.69	2.74	33.05

[a] Values quoted are the mean of the absolute monthly percentage changes of each index
[b] Values are reported in percent
[c] *Yorkshire H* refers to the *Yorkshire and the Humber* region

6.4 Additional stratification through property type

The Price Paid Data exposes another variable on which we can perform stratification: property type. Through the use of geohash$^+$ (see: Section 5.3.1), we can encode the property type as an additional variable by prefixing it to the geohash string for each property. Thus, each geohash will now begin with a *D*, *S*, *T* or *F*, according to their respective type.

Through this encoding method, the GeoTree will only return properties which share the same property type prefix as the property being searched for, within a small area of the home in question.

6.4.1 National index

6.4.1.1 Time series analysis

Figure 6.15 demonstrates our *GeoPrice* index with additional property type stratification against the ONS house price index. The index generated by the *GeoPrice* algorithm is very similar to the original result, which did not include property type matching. This seems to suggest that the addition of this supplementary stratification parameter does not add a great deal of value to the index.

On deeper thought, this seems rational. Consider that the first variant of the algorithm presented was using geospatial proximity alone to perform the stratification process. It would typically be the case that properties of the same type are co-located. For example, an apartment is always part of a block, terraced houses always come in a row and detached houses are usually secluded, with the nearest neighbours also having detached properties. The lone exception to this may be the case of a semi-detached house sharing a border with a terraced house. However, in these cases, the value of the properties typically trend together, as the valuations will be very similar, with the semi-detached home attracting a small premium. Thus, even in this instance, it is likely that excluding the terraced neighbours does not significantly reduce model noise.

These results add credence to our argument of why geospatial stratification works (see: Section 3.2.2.2); the strong autocorrelation effect exhibited by housing seems to be such that our model is capturing the explanatory power of property type on the

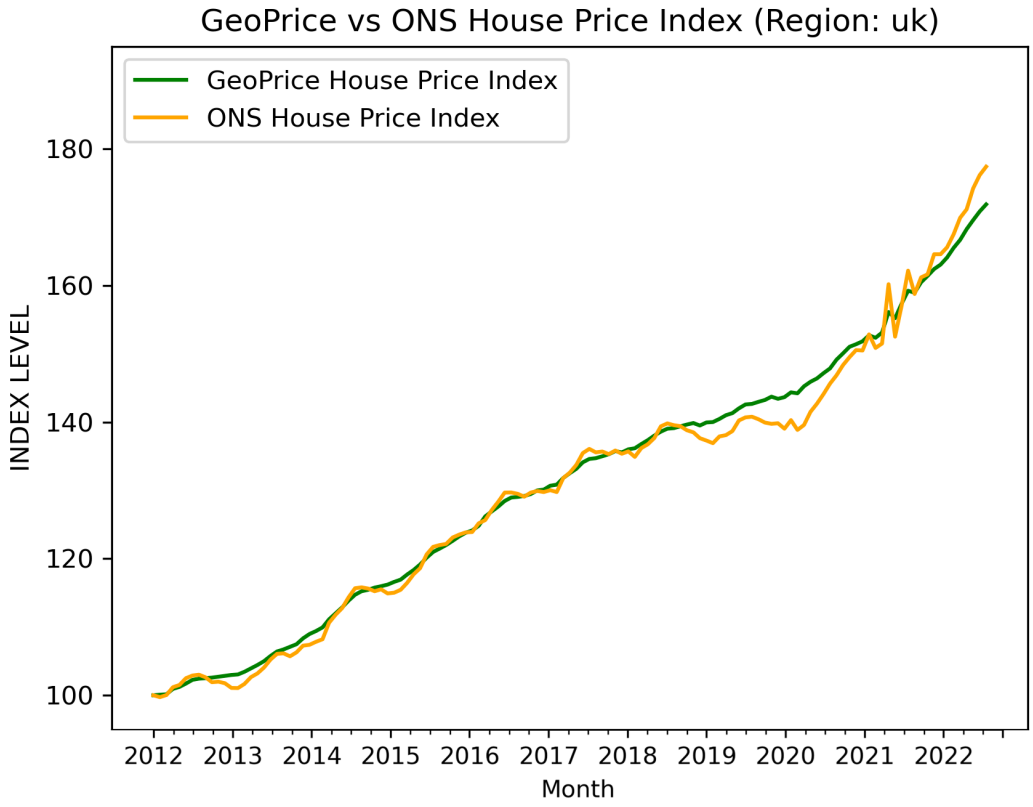


FIGURE 6.15: ONS vs *GeoPrice* (w/property type) House Price Index [UK] from 01-2012 to 09-2022

value through geospatial autocorrelation alone, without the need to explicitly know the property type. This is a highly encouraging finding in terms of the confidence behind our model's core conceptual design.

TABLE 6.10: Smoothness Metrics for ONS and *GeoPrice* Indices [UK]

Model	Mean[a]	Median[a]	St. Dev[b]	Min[b]	Max[b]	St. Dev of Diffs	MSM
GeoPrice [c]	0.45	0.42	0.34	-0.58	1.96	**0.45**	**0.87**
ONS HPI	0.75	0.54	1.0	-4.79	5.73	1.58	11.24

[a] Values quoted are the mean of the absolute monthly percentage changes of each index
[b] Values are reported in percent
[c] *GeoPrice* algorithm is run with geospatial and property type stratification

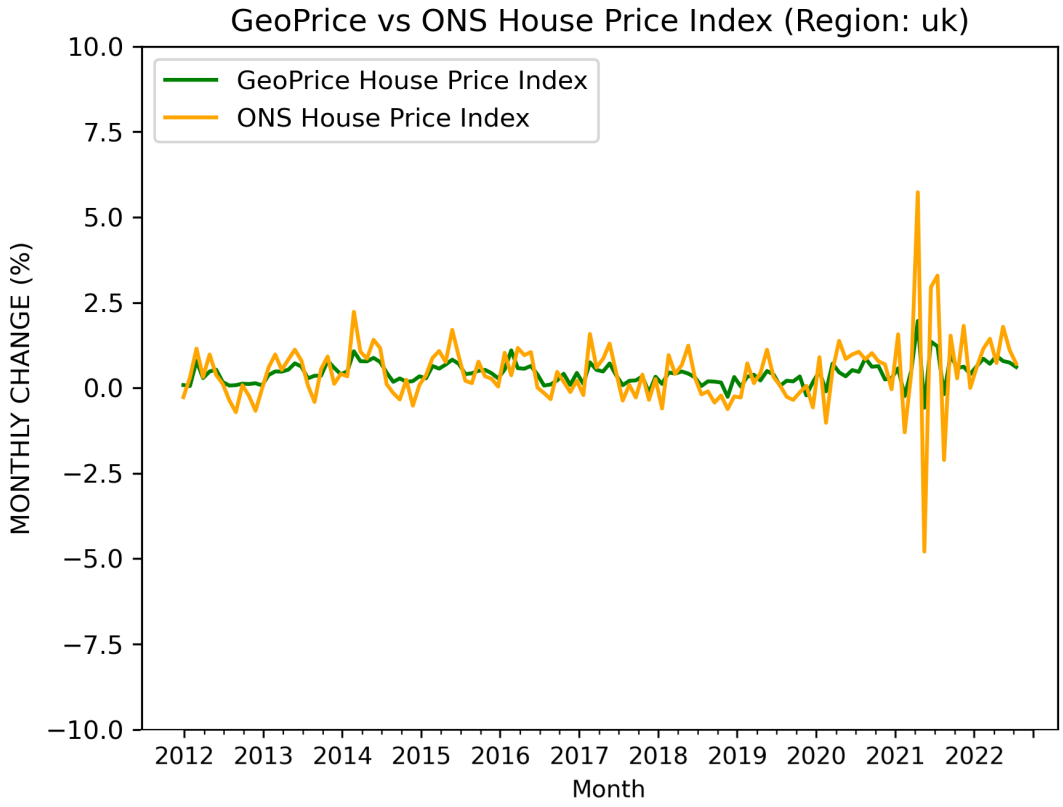


FIGURE 6.16: ONS vs *GeoPrice* (w/property type) House Price Monthly Change (%) [UK] from 01-2012 to 09-2022

6.4.1.2 Smoothness metrics

The smoothness metrics in Table 6.10 corroborate our prior findings. Neither the standard deviation of the differences nor the mean spike magnitude are improved by the addition of property type stratification. Despite this, the model retains comparable performance with the geospatial-only variant of the index, with negligible differences across all of the statistics shown.

In terms of correlation with the ONS hedonic regression model, Pearson's r coefficient is marginally higher at **0.8644**.

6.4.2 Regional sub-indices

Similarly to the results for the national index, regional sub-indices are not significantly changed through the introduction of property type as an additional, explicit attribute. The index level and monthly change results for the regional sub-indices with property type included can be found in Section A.5.

Smoothness metrics for the regional sub-indices where the property type has been encoded as an explicit variable are shown in Table A.1.

6.4.3 Property type sub-indices

A tangible benefit of having the property type variable as an explicit value is the ability for use to construct sub-indices for each of the property types. While, as discussed, it did not add any additional smoothness to our aggregated indices, it does allow us to filter our strata and run the *GeoPrice* algorithm on each subset.

Figure 6.17 demonstrates the sub-indices for each distinct property type, superimposed with the ONS house price index for the equivalent property type. The trend of all sub-indices is very similar between both models, across the majority of the sample period.

Recently, the detached home index has started to drift further from the corresponding *GeoPrice* index. It is interesting to note that the transaction volume for detached homes has substantially reduced versus the historical norm over the same time period that this divergence has been observed, as shown in Figure 6.5. Perhaps this change in the transaction sample has caused increased volatility and the surge

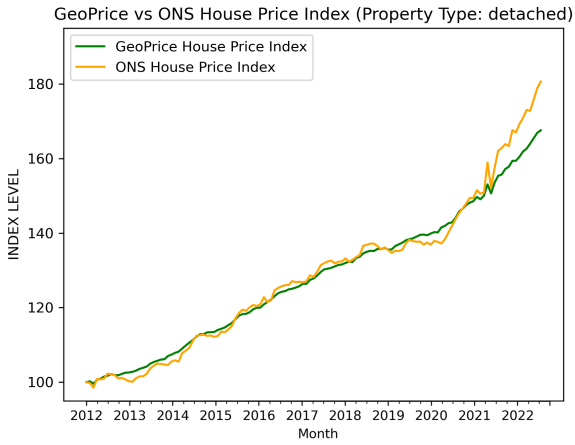
in value for this particular type of property recently, in the ONS hedonic regression model. This disparity is likely another driver behind the previously discussed drift in the national index across the same time period.

Table 6.11 demonstrates the robustness of the *GeoPrice* house price index across every property type. The standard deviation of the differences has been reduced by at least 59.75% in all four sub-indices versus the ONS house price index. The mean spike magnitude, on the other hand, is over 84.8% lower in each case. Interestingly, the two sub-indices with the most volatile performance in the ONS model, terraced houses and apartments, are the least noisy sub-indices in the *GeoPrice* algorithm. This coincides with the correlation results; the terraced homes and apartments monthly change values have a Pearson's r coefficient of approximately 0.6. On the other hand, the detached home sub-index has a correlation of 0.8 with the ONS equivalent, while the semi-detached index has the greatest similarity, at 0.84.

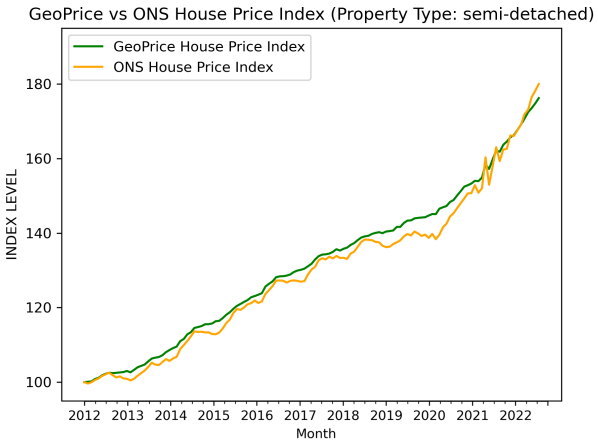
TABLE 6.11: Smoothness Metrics for ONS and *GeoPrice* Indices, on each property type [UK]

Model	Mean[a]	Median[a]	St. Dev[b]	Min[b]	Max[b]	St. Dev of Diffs	MSM
Detached / *GeoPrice*	0.46	0.36	0.41	-1.57	2.0	**0.64**	**1.88**
Detached / ONS HPI	0.76	0.5	1.02	-4.29	5.35	1.59	12.36
Apartment / *GeoPrice*	0.44	0.43	0.39	-0.91	1.63	**0.57**	**1.31**
Apartment / ONS HPI	0.91	0.78	1.12	-3.1	4.8	1.8	14.15
Semi-detached / *GeoPrice*	0.47	0.39	0.4	-0.74	2.26	**0.57**	**1.45**
Semi-detached / ONS HPI	0.77	0.6	1.0	-4.58	5.45	1.59	10.63
Terraced / *GeoPrice*	0.48	0.48	0.36	-0.45	1.72	**0.46**	**0.92**
Terraced / ONS HPI	0.91	0.64	1.28	-6.52	6.98	2.06	19.09

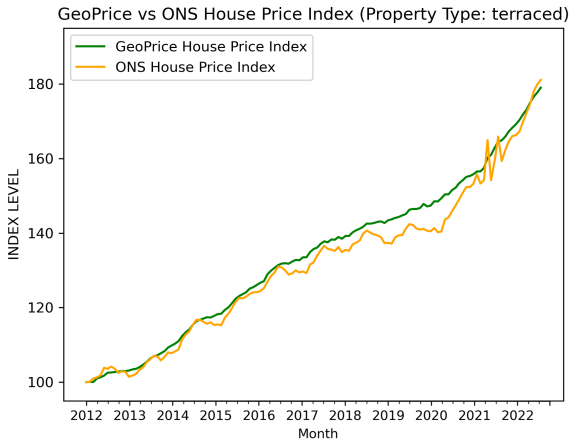
[a] Values quoted are the mean of the absolute monthly percentage changes of each index
[b] Values are reported in percent



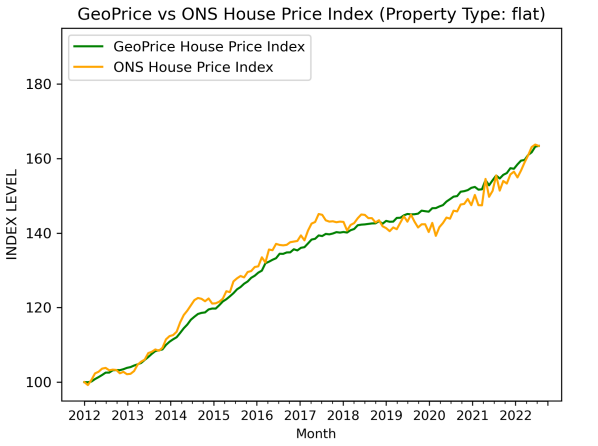
(A) Detached home indices



(B) Semi-detached home indices

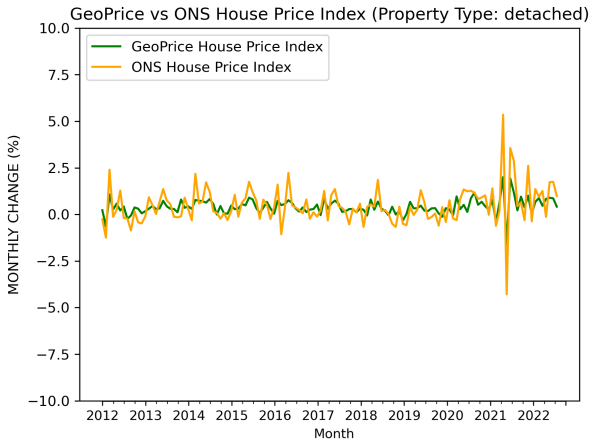


(C) Terraced home indices

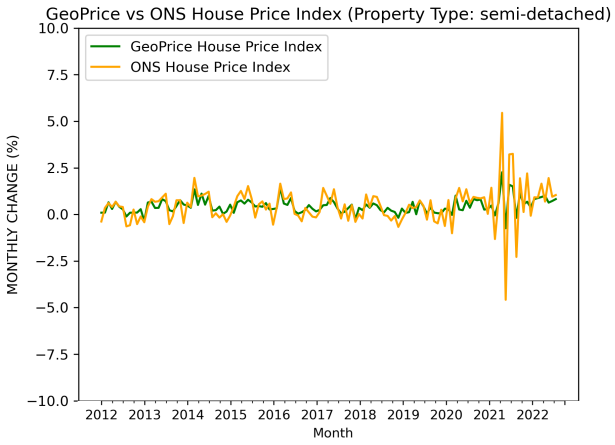


(D) Apartment indices

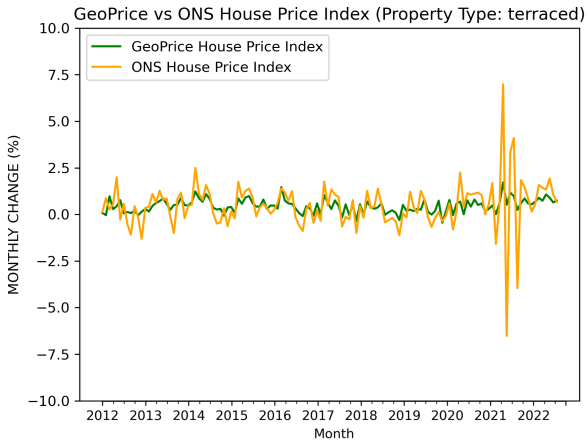
FIGURE 6.17: ONS vs *GeoPrice* House Price Index [UK] from 01-2012 to 09-2022, per property type



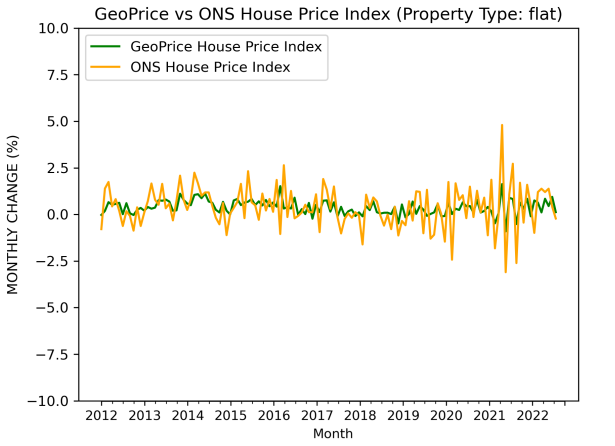
(A) Detached home monthly changes



(B) Semi-detached home monthly changes



(C) Terraced home monthly changes



(D) Apartment monthly changes

FIGURE 6.18: ONS vs *GeoPrice* House Price Monthly Change (%) [UK] from 01-2012 to 09-2022, per property type

6.5 Chapter Summary

The methodological tweak of considering the entire distribution of neighbours near a given property, rather than just the median neighbour, clearly pays dividends on application to a large dataset, such as the UK's Price Paid Data. Indeed, the *GeoPrice* model achieved only a modest improvement in smoothness on the Irish market (relative to the RPPI), the results of applying the enhanced algorithm to the UK market are a marked improvement. The *GeoPrice* index maintains a high level of correlation with the national hedonic regression model for the UK, while achieving a considerably greater degree of smoothness.

Furthermore, the algorithm is entirely automated and can be recomputed rapidly and at high frequency, owing to the performance benefit delivered through use of the *GeoTree*. With this data structure in hand, it is possible to recompute the index within one hour of a new release of the Price Paid Data. The ONS hedonic regression model, on the other hand, is released on a lag of several weeks from the publication schedule of the Price Paid Data and requires substantial human effort to produce on a monthly.

These results extend to the various sub-indices; both regional and by property type. Once again, the *GeoPrice* model produces a very similar trend over the long term to the ONS house price index, while delivering near an order of magnitude less noise. Moreover, the *GeoPrice* model realises these advances without the introduction of any of the additional attribute data or neighbourhood quality data which the ONS hedonic regression invokes. Our index is based solely on public facing sale transaction data and thus could be reproduced by any interested party, at will.

Chapter 7

Conclusion

To conclude this thesis, the original key goals and objectives of the *GeoPrice* model will be discussed, with an evaluation of the extent to which those aims were achieved. Furthermore, potential applications of the model, along with future research which could be undertaken to advance the methodology further will be outlined.

7.1 Analysis of objectives

Section 1.5 outlined the primary objectives of this research project, with List 1 indicating the desired attributes and features of the property price index methodology being proposed. The overarching goal of the work was to introduce an efficient, scalable and flexible property price index model, which could be utilised by a wide number of stakeholders in the property market. As discussed in Chapter 2, it is challenging for market participants, aside from perhaps governments and central banks, to gain access to alternative sources of information on the housing market, or verify the statistical veracity of the *official* models produced by the national statistical offices.

In many instances, these models, often produced through a conventional hedonic regression, are the only generally available metric on the state of the house price index for a given region. The concept of other econometric areas of interest such as inflation, economic strength or labour market dynamics only being measurable via a single data release produced by a single party seems absurd. Similarly, it would seem irrational to measure the trend of any other asset class, such as bonds or equities via a single, irreproducible metric. Thus, when one compares the housing market to any other econometric data release or large asset class, this makes the housing

a highly unusual and isolated case.

The house price index model introduced in this thesis has been created with the goal of solving these issues. The core conceptual framework of the model is simple to understand and based on the demonstrable geospatial auto-correlation exhibited naturally by housing. Thus, a large amount of expertise is not needed to understand and utilise the proposed methodology. Furthermore, the thesis has demonstrated the ability for the model to deliver not only equivalent, but superior results to hedonic regression models. In each use-case of the *GeoPrice* index explored in the thesis, this has been achieved despite the use of a significantly more sparse and frugal dataset, including publicly available datasets for the United Kingdom and the Republic of Ireland, respectively.

Key Goal 1: The model must be fully automated

The first goal outlined in List 1 asserted that the *GeoPrice* model needed to be capable of ingesting a dataset of property transactions (or asking prices) and compute the house price index without any human intervention, i.e. the index must be completely automated.

In each of the applications of the model demonstrated throughout Chapter 4, Chapter 5 and Chapter 6, the model needed only to be fed with an updated dataset and the *GeoPrice* index could be reproduced at will, without any human involvement in the process. Indeed, all that would be necessary to generate a fully automated model which automatically updates upon new data releases is to build a web-scraping script which pulls the raw transaction data from the public source and saves it into any suitable format (e.g. file, database, etc.) and to setup a scheduled job to run the *GeoPrice* index as frequently as desired.

The importance of this goal is two-fold. Firstly, as discussed as length in Chapter 2, many market stakeholders do not have the resources nor the expertise to regularly produce a house price indicates which requires heavy domain knowledge, human oversight and regular maintenance. Introducing an entirely automated model reduces the barrier to entry, allowing any interested parties with a dataset in hand to fit an accurate, performant house price index to that data with low resource cost.

Secondly, a notable drawback of the conventional hedonic regression models produced by national statistics offices is the lack of transparency and reproducibility around their house price index models, as discussed in Chapter 1. If the *GeoPrice* model were to require human expertise and intervention in the production of the output index, this would defeat the purpose of making house price index methodology more accessible and corroborable. The transparency gained by introducing an index which can operate on publicly available data would be lost if the methodology is not straightforward to the layperson.

Key Goal 2: The model must be capable of operating on a minimal set of attributes

The second objective of the *GeoPrice* index was to ensure that the model required nothing more than a minimal set of data on each transaction in order to produce an accurate index, namely: the sale date, the sale price and a set of GPS co-ordinates (or an address which could be geocoded to such). This was a relevant goal as most publicly available property transaction datasets, such as the *Property Price Register* used in the initial application of the model in Chapter 3 include no data on any characteristics of the property.

An index methodology which requires more than these attributes to achieve a robust index breaches the goal to deliver a model which is more accessible to market stakeholders, as sourcing this additional data is typically time-consuming, expensive and oftentimes outright impossible for the majority of stakeholders. Furthermore, the *GeoPrice* was designed to leverage the spatially auto-correlated nature of housing; if it were not possible to derive a usable index from this factor alone, the basis of the model would be somewhat flawed.

Both Chapter 3 and Chapter 4 demonstrate that the *GeoPrice* index is capable of outperforming a conventional hedonic regression model while only using a minimal set of explanatory data on the characteristics of the property. Furthermore, the analysis in Chapter 6 indicates that this minimal set of attribute data was sufficient to significantly outperform the ONS' widely used hedonic regression house price index, which requires a plethora of exogenous attribute variables from multiple private data sources.

Key Goal 3: The model should be flexible enough to operate on different types of datasets

Another key aim of the *GeoPrice* index was to ensure methodological flexibility. In other words, the model should be flexible enough to be capable of producing a house price index on a variety of different kinds of datasets, e.g. transacted prices, asking prices etc. As discussed at length in Chapter 3, various market stakeholders may have distinct property datasets, including both bespoke and public data, available to them and thus it is critical that the methodology is pliable enough to cater for these varied cases.

Chapter 5 demonstrated the results of applying the index to a dataset of Irish listed asking prices, which was compared against the results of Chapter 3, where the model was fit on real sale transactions in the same region. The total number of samples and timespans of each of these datasets differed from one-another, however, the *GeoPrice* model was capable of producing an informative and accurate model on each of them.

This flexibility is not a given with all house price index methodologies. For example, the repeat sales methodology outlined in Chapter 3 would not be amenable to an asking price dataset in the majority of cases. Typically, asking price data is only available over the time a particular online property portal is operating. Furthermore, property portals are chosen at the discretion of the seller; a given portal is not guaranteed to have every house on the market listed on it.

As a repeat sales model relies on comparing multiple sales of precisely the same property, it is unlikely that a sufficient number of houses would i) appear multiple times within the lifespan of a particular portal and ii) have multiple successive homeowners select the same portal to host the listing. For these reasons, applying this particular methodology to asking prices is challenging and often times infeasible.

Given that the *GeoPrice* model could be considered, in some ways, to be a more generalised evolution of the repeat sales model, whereby each particular property is matched with a comparable record in each preceding sale period, it does not suffer from this same limitation and does not require the extensively lengthy history of

samples which a repeat sales model does. These characteristics of the model make it suitable for application to asking prices.

These applications have illustrated the potential for our model to generalise to distinct markets in different countries, as well as the capability of modelling asking prices, as well as transacted prices. These findings offer substantial evidence that we have achieved our goal of introducing an accessible and adaptable model.

Key Goal 4: The model must be capable of incorporating additional attribute data, if it is available

While the *GeoPrice* model has been designed with a conservative and frugal dataset in mind, retaining the optionality to factor in additional property attributes was an important aspiration. In some potential applications of the *GeoPrice* model, data which would assist in further improving the accuracy of the comparable matching process is likely to be available to the user. It would be foolish if this data was not incorporated into the model and used to attempt to boost the performance.

Indeed, Chapter 5 demonstrated that the inclusion of number of bedrooms alongside the geospatial matching offered a significant increase in the smoothness of the output index. This was achieved with minimal additional complexity and no loss in model efficiency. This is not a surprising outcome; as discussed in Chapter 1, many of the factors impacting the valuation of a given property will be shared by those in its proximity. However, while there is some likelihood that neighbours may often share a similar number of bedrooms, there will be many exceptions and violations of this rule. As such, one would expect that combining the number of bedrooms with the geospatial matching should deliver noteworthy performance gains; something which was proven to the be the case.

By contrast, in Chapter 6, the *GeoPrice* model was armed with the property type of each sale transaction, which was again used alongside the geospatial proximity matching process. However, in this particular instance, the performance gains were near negligible; the geospatial matching process alone appeared to be sufficient to infer the impact of this attribute. Upon thought, this is not a particularly surprising result; it is highly typical for properties of the same type to be located near to one another. Apartments, for example, will always come as part of a block.

In both of these cases, the *GeoPrice* index was capable of ingesting additional data on the characteristics of the properties being fed to the model. In the former case, this additional context was leveraged to improve the performance by a considerable amount. In the latter case, where the attribute in question was mostly inferrable through the spatial auto-correlation effect, the data did not significantly improve the model, but also did not cause any notable distortion nor did it hinder the index in any manner.

Key Goal 5: The model should be scalable and performant enough to operate on datasets of different sizes

Chapter 4 illustrated some of the challenges in performing geospatial proximity searches and geospatial clustering on large datasets. Indeed, the performance bottlenecks of the initially proposed *GeoPrice* methodology in Chapter 3 rendered it unsuitable for expansion to larger datasets with a greater volume of transactions, as the execution time grew quadratic with the number of the samples.

The introduction of the *GeoTree* in Chapter 4 succeeded in removing this bottleneck, through speeding up the efficiency of the model computation by multiple orders of magnitude. This was achieved with a negligible reduction in model accuracy when compared against the initial formulation of the index. Furthermore, the scalability of the *GeoTree* was demonstrated, showing the reduction from quadratic complexity to linear complexity and, as a result, achieving the stated performance objective.

Chapter 6 leveraged this scalability by applying the *GeoPrice* model to a dataset with monthly transaction volumes around twenty times larger than the original index analysis. Despite the large increase in dataset size, the *GeoPrice* model retained the capability to compute the index from start to finish within fifteen minutes.

Despite these variations in data sample size, the model demonstrated its ability to produce a high-quality, accurate index on all applications. This was further established in Chapter 6, where sub-indices were generated for each region in England and Wales. The *GeoPrice* model significantly outperformed the smoothness and the robustness of the ONS hedonic regression model on all of these regions, proving its ability to deliver an accurate, scalable and efficient house price index regardless of

the input sample size. Moreover, this is achieved while maintaining high levels of month-on-month correlation to the benchmark, indicating similar behaviour is being captured, but with materially less noise contamination in our proposed model.

Key Goal 6: The index should be capable of updating rapidly once new transaction data becomes available

As discussed previously, the delivery of scalability objective has resulted in the *GeoPrice* index being capable of computing within a matter of minutes, even on large datasets such as the one used in Chapter 6. As such, there are no bottlenecks remaining which would prevent the index rapidly updating upon being fed with new transaction data, bar the availability of the data itself.

Thus, the *GeoPrice* model is capable of automatically updating, on a schedule, within hours of a data release from any source of regularly updating property valuation data, be it asking prices or sale prices. The delivery of this goal supports the aim to reduce the lead time from the data becoming available, to house price indices based on this data being published. Typically, this lead time is in the order of months for popular conventional models, leading to the propagation of stale information, which is less conducive to an efficient market.

In order to more explicitly demonstrate that this goal has been met, later in this chapter a proof-of-concept automated house price index web-platform will be presented, operating on both of the transaction datasets introduced in Chapter 3 and Chapter 6, respectively.

Key Goal 7: Despite the restrictions outlined, the minimal index must deliver equal or better performance than conventionally used models on rich datasets

In each application of the *GeoPrice* model where a conventional benchmark was available, the ability for the index to outperform in both accuracy and smoothness, while retaining high levels of correlation, was proven. While this admittedly ranged from a slight improvement in Chapter 3 to a substantial improvement in Chapter 6,

in both cases, it was achieved under a significantly more restrictive set of conditions than the benchmark was subjected to.

In both cases, the conventional benchmark hedonic regression models were equipped with a range of explanatory attribute variables, a team of expert statisticians and ample lead time in the order of several weeks from the release of the transactions to the publication of the model results. By contrast, the *GeoPrice* index was given a minimal set of data on each transaction, the restriction of complete automation and the goal of publishing the index results on the same day that the transactions became available.

As such, it has been demonstrated that the *GeoPrice* model has met the most critical objective of outperforming the benchmark models in accuracy, while being subjected to significantly harsher conditions, concluding the final of the research goals which were set out in List 1.

7.2 Thesis contributions

The thesis objectives of building credible evidence behind the viability of a novel method of stratifying property price transactions have been addressed thoroughly, indicating probable merit and convincing advantages in adopting this methodology as a tool for measuring and monitoring the housing market generally. The analysis undertaken has identified a number of key findings in the field of property price research.

Key Finding 1: Spatial auto-correlation alone is sufficient to derive an accurate house price index

Perhaps the most vital conclusion of the research undertaken is the validation of the claim that the spatially auto-correlated nature of the housing market alone is enough to produce a performant house price index model.

An unsubstantiated claim had earlier been made by O'Hanlon, 2011 that the Property Price Register dataset used in Chapter 3, containing sale transactions for

the Republic of Ireland, was an "impractical" dataset for use in house price modelling, owing to the lack of any attribute data. They concluded that the dataset offered no viable method of mix-adjusted and thus could not be used to generate an accurate house price index.

As demonstrated through Chapter 3 and Chapter 4, this claim is evidently false. Not only has this dataset, in tandem with the *GeoPrice* model, been proven to be sufficient for the purpose of creating a performant house price index; the resulting house price index outperforms the hedonic regression model formalised and adopted by O'Hanlon, 2011 and the Central Statistics Office of Ireland in the original research piece where said claim had been made.

The basis of the *GeoPrice* model was further bolstered in Chapter 6, where the index generated purely through the premise of spatially auto-correlated matching was again capable of outperforming a widely-used, popular hedonic regression model. Furthermore, as discussed prior, the addition of the property type attribute did not make any significant difference to index performance; indicating once again that spatial auto-correlation was capable of inferring the vast majority of the explanatory power that this variable has on property value.

These findings suggest that the widely observed, yet under-leveraged impact of spatial auto-correlation on properties can be a powerful tool in analysis of the housing market and the potential to garner insight from frugal property datasets should not be dismissed without thorough investigation; it is not necessary to have a rich, perfect dataset of attributes in order to achieve compelling and useful results.

Key Finding 2: Further research should be undertaken on novel methodologies beyond bolstering the shortcomings of existing, conventionally used models

As discussed in Chapter 1, a great deal of research has been done on attempting to address the most significant drawbacks of hedonic regression models; one of which is the frequent lack of accounting for geospatial effects in the model specification. Recent additions to the literature have explored methods of modifying and augmenting the hedonic model formulation to better incorporate these geospatial effects, which has shown encouraging benefits and improvements to model accuracy.

A key finding of this thesis has been to illustrate that hedonic regression models should not necessarily be the default go-to model for all house price index analysis. These is demonstrable merit in investigating and formulating alternative methodologies, such as the *GeoPrice* index introduced in this thesis. Not only has this novel model indicated that it is possible to outperform conventional hedonic regression models through leveraging their "greatest weakness", according to O'Hanlon, 2011, but this has also been achieved with a publicly accessible dataset and methodology, and no publication lag nor lead time.

Further research should be undertaken in investigating novel methods of measuring house price trend, particularly those which are capable of incorporate spatial auto-correlation, which has been verified to hold strong explanatory power over housing valuations. The value and potential of automated, performant, timely and transparent methodology over that which is opaque, lagged and resource-intensive should not be underestimated.

Key Finding 3: If sufficient, timely input data were to be available, it would be possible to compute a near real-time house price index

As has been discussed at length throughout the thesis, one of the problematic limitations of existing, conventional house price indices is the significant delay with which they are published. Furthermore, given that they generally cannot be reproduced by third parties due to lack of access to the input data, it is often not possible for stakeholders to produce a more timely model privately.

Given the importance of house price indices across a wide range of use cases, as discussed in Chapter 2, this lead time to publication should be a significant concern. The results of these models are used in monetary policy, fiscal, lending and planning decisions, among others; some of the most critical decisions in terms of influence over the broader economy. The fact that these models are currently being used with a lag of typically around two months means that stale, out-of-date information is being factored into said key economic decisions.

With the research presented in this thesis, one of the findings which potentially has the most broad-reaching impact is the proof-of-concept demonstration that it *would* be possible to compute a near real-time house price index, if sufficient raw

input transaction data were to be made available to the model in real-time. The final formulation of the *GeoPrice* index presented in Chapter 6 illustrates that a house price index can be produced and published end-to-end within a matter of fifteen minutes on a large region.

Aside from the saved resources that this rapid, automated calculation offers, having a more timely and up-to-date index available when making policy decisions would likely be of great use to central banks and governments. It is not unusual for these parties to seek faster measures of other economic indicators, such as inflation and labour market dynamics, to get a more reactive measure of market changes than that provided by what is considered to be their benchmark or gold-standard metric.

While transaction data availability still remains something of a bottleneck at present, the several week delay typically added by conventional index calculation itself has been removed entirely by the *GeoPrice* model. Furthermore, Chapter 5 demonstrated that the model is also capable of being applied directly to listed asking prices. These would be available from multiple sources as a real-time dataset, leading to the potential of launching a real-time house price index, albeit one with the limitations of using asking prices rather than actual sales, which result in different, yet correlated econometric measures.

Key Finding 4: Some attribute data remains relevant and performance-enhancing, even when spatial auto-correlation is accounted for

Despite the clear evidence that spatial auto-correlation is a sufficiently explanatory feature in order to outperform conventional hedonic regression models, our findings in Chapter 5 demonstrate that there is still value in incorporating some attribute data into price index construction, if said data is available to the end-user.

In the application of the *GeoPrice* model to a dataset of asking prices in Chapter 5, two variants of the index were produced; one with geospatial matching alone, the other with a combination of number of bedrooms and geospatial matching. The results confirmed that the inclusion of the number of bedrooms offered a significant boost to the smoothness of the output model, while remaining highly correlated. As discussed previously, this is not a surprising outcome, given that the number of

bedrooms would be one of the attributes which is more likely to differ among neighbouring properties and thus spatial auto-correlation would be expected to have only partial explanatory power over this particular price-impacting variable.

This finding justifies the inclusion of the flexibility goal outlined in List 1 and Section 7.1, which was aimed to address such scenarios. Although the primary aim of the goal was to introduce a new type of house price model which could function on minimal public data, the capability to retain compatibility with richer datasets was crucial desired functionality. These findings suggest that further research should be undertaken to combine geospatial matching methods alongside the property attributes which are less spatially auto-correlated, which may result in further performance enhancements.

7.3 Uses of the *GeoPrice* model

As discussed in Section 2.4, a large number of stakeholders hold an interest in the property market. Any one of these categories of stakeholders could employ our proposed model as a method of analysing property prices. Governments and central banks, as discussed, rely heavily on housing market statistics and analysis for their policy and budgetary decisions. The interest rate set by the central bank has a high correlation with mortgage interest rates, resulting in the need to carefully balance this with households' affordability. These parties are currently working with stale, slowly updating data in taking their views on the market, which is a significant issue when considering the impact of the decisions being made. It is not necessarily the case that this proposal must serve as a full replacement for existing, long-trusted methodologies, however, it may have meaningful use in providing a more up-to-date view than the formerly mentioned tool, given the lower lead time.

Letting agents and property portals would likely also have a keen interest in alternative property price index measurements. Indeed, these stakeholders often have large databases of listed properties and their respective asking prices, in addition to further attributes which could be explored for deeper stratification, as we explored

in Chapter 5. Through use of their own asking price datasets, these market participants could generate bespoke property price indices for their clients, better informing them of local trends in the property market. Furthermore, given their dataset is being constantly updated with new listings, the potential exists for them to roll this out as a novel daily updating service, which may help keep homeowners and prospective buyers logging on to their platform to check the latest trends; boosting engagement metrics and improving their sell-through rate on listings.

Finally, lenders are another party who may have a strong interest in producing their own, custom house price index. Mortgage lenders take on a great deal of risk, given that a house is the most expensive asset owned by most typical citizens. This risk fluctuates in magnitude according to the housing market behaviour, economic environment and interest rate set by the central bank. If these lenders could produce a more up-to-date house price index with their own database of mortgage returns completed by their customers, this may give greatly enhanced visibility on the market and better inform their lending practices. For example, rather than waiting two months delay for the official house price index, lenders could spot a potential turn in the market one day after the data is collected by running their own model and take action to reduce their risk on upcoming approvals and the interest rates being set.

As such, there are many potential users of the *GeoPrice* house price index and possibly many more which we have not mentioned. Due to the enormous size of the housing market, there are sure to be countless observers taking a keen interest in this asset class and thus the research conducted in this thesis is conceivably relevant to any of them.

One potential drawback which remains in the proposal is that it relies on GPS co-ordinates or, more specifically, geohash encodings of GPS co-ordinates in order to perform the voting and stratification periods. While these were not readily available on a property-by-property basis in Ireland, we were able to source them through geocoding API services. In our analysis of the UK data, GPS co-ordinates were made freely and publicly available by the ONS.

If the model is to be applied to any market, the user must ensure the ability

to obtain these from the home addresses. Nevertheless, these are typically significantly easier to acquire than the detailed property characteristics required by hedonic regression, owing to the prominence of automated geocoding and mapping APIs available on the open market. As such, the barrier to entry for the model remains orders of magnitude lower than that of a conventional hedonic regression model and this limitation can typically be easily worked around.

7.4 Future work

While the thesis has succeeded in meeting the initially outlined research objectives and desired functionality for the *GeoPrice* house price index model, the findings presented leave the door open to a great deal of prospective future work which could be undertaken. These potential directions include both practical applications of the model and more theoretical avenues of analysis which may result in further improvements and evolutions of the *GeoPrice* methodology.

7.4.1 Application-based extensions

There are many potential practical applications of the model, which would serve to demonstrate the use-cases and contribution of the research. In order to further validate the flexibility and adaptability of the model, it would be desirable to continue to benchmark the model in different regions versus their official national counterparts.

Automated *GeoPrice* indices could be built for countries across the world, including the United States, European Union nations and countries in Asia and Oceania. Each of these countries will likely have different standards and formats of available sale transaction data, some of which will likely be private or charged, while others will be open, public datasets. Morphing all of these datasets into a clean, unified format for ingestion by the *GeoPrice* model would be the most time consuming component of this task.

The target vision for the *GeoPrice* model would be to launch an online web-platform which is free and public facing. This platform would offer *GeoPrice*-powered indices for a palette of countries and regions across the world, allowing for different levels of regional granularity in different countries. Furthermore, a suite of tools

could be made available to users which perform analysis of cross-nation correlations and patterns in property prices, while also indicating the countries which have a strongly trend housing market.

The platform would be designed to update automatically once new data for each country becomes available; scrapers or feeds would be set up to automatically poll, pull, clean and store data from the relevant source in a single, unified database format. Once this has completed, a signal could be sent for the index to re-compute for that particular region, with the results becoming available for users within a matter of minutes.

It would also be possible to connect different property price data feeds to the platform for each regions where they are available, which could be transactions, mortgage data, asking prices, etc., depending on what is available in the region in question. This project would be an ambitious undertaking to build and maintain, however, the fundamental methodology to support it has been laid out in this thesis and the *GeoPrice* index should be robust, accurate and performant enough to enable the creation of such a resource in its current state.

A basic proof-of-concept demo of this application has been launched, in order to demonstrate the practical use of the work undertaken in this thesis. The web platform contains an automatically updating house price index for the Ireland and the UK, based on the same two datasets explored in Chapter 3 and Chapter 6. Each month, once new data becomes available, a scheduled job pulls the data and recomputes the *GeoPrice* indices, publishing their output to the platform for the user to view. This is achieved with little-to-no lag and no human intervention, rendering the system completely automated.

Furthermore, the tool illustrates the ability of the model to generate sub-indices for smaller localities within the region in question, as shown in Figure 7.1. Using the interface, users can drag and select a specific area on the map and a bespoke index will immediately be generated and displayed, focusing on the price trends within the area they selected. While this is a basic prototype of the more advanced platform described previous, it nonetheless demonstrates the viability of such a tool being developed and expanded with additional regions and analytical tools.

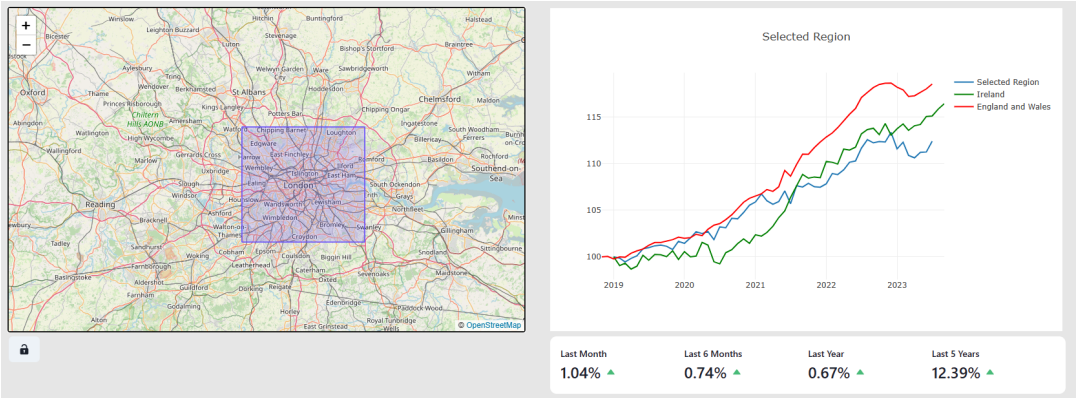


FIGURE 7.1: Proof of concept *GeoPrice* web platform demo

7.4.2 Theoretical advances

A number of potential theoretical advances could be explored, in order to build upon the foundations set by the *GeoPrice* model. These range from ideas for methodological enhancements to improve the performance of the index, to alternative formulations of the core conceptual framework which could allow the application of the idea to a distinct use case. Below, some suggestions for further research projects involving the *GeoPrice* model will be briefly outlined.

Enhancing how attribute data is incorporated into the model

In Chapter 5, the thesis explored how additional attribute data (the number of bedrooms, in that specific example) could be incorporated into the model specification in order to improve the accruacy of the comparable matching process. While this was successful in significantly boosting the performance of the model, there are potential limitations of the current specification, if pushed to an extreme use-case.

For example, if a user decided to use several attributes as part of the matching process, it is possible that the number of strata would then become so large as to result in time periods where there is no proximate neighbour in the sample. For example, it is not guaranteed that there will be a six-bedroom, three-bathroom, detached property with a floor area range of X - Y square meters which is located in a reasonable vicinity of a transacted property in every prior month over which comparable sample is being drawn. While this is an extreme case involving a very large number of attributes in addition to somewhat of an outlier on the upper-end of

the property distribution, it demonstrates a limitation of the matching system and a drawback of mix-adjusted models in general, as discussed in Chapter 2.

Research could be undertaken into using the results from similar strata to improve the matching accuracy and sample size in cases such as these. For example, the model could intelligently look at slight variations to the matching criteria; five-bedroom and seven-bedroom neighbours could also be incorporated, with an adjustment factor applied to their prices based on a longer term analysis of the difference between those strata and the optimal stratum.

This could also be extended to not only be applied to attributes, but to the geospatial component as well. While the *GeoPrice* currently looks at a collection of nearby neighbouring properties sold in prior time periods, it would be interesting to explore the effect of taking a weighted average of larger and larger neighbourhood vicinities. For example, the price comparison could be based 70% on neighbours within a few blocks, 20% on the next *GeoTree* level up from that, and the final 10% on another level up from that. This experiment could reveal some interesting insights into how far the spatial auto-correlation effect reaches and whether there is an identifiable point at which the performance begins to deteriorate.

Combining transactions and asking prices into a single model to add forecasting abilities

As demonstrated in the thesis, the *GeoPrice* model is capable of generating an index on both transacted properties and property listings. An evolution of the model which could potentially be of great use to market stakeholders would be to attempt to combine this two distinct sources into a single model, in order to forecast the upcoming movements in property prices.

Very few robust forward-looking indicators of property price trends exist in the industry. The reasoning for this is likely the same reason that alternative, privately produced house price index models are rarely seen; acquiring the data and fitting a conventional model to is challenging and resource-intensive. However, with the more straightforward, frugal and automated methodology offered by the *GeoPrice* index, there is renewed potential to conduct more exploratory analysis on creating forecasts for housing through the combination of sale transactions and asking prices.

As listed asking prices are a leading indicator of upcoming sales, it should theoretically be possible to gain a forecasting advantage through their application. Almost all sales which eventually appear in the transacted property dataset will have been listed for sale many months prior on a property listing platform of some description. While it is possible for the prices to change based on negotiations, the trend of the initial listed price and any updates made to it by the seller after listing should be sufficient to offer some predictive power over the upcoming move in the market.

Analysis would need to be undertaken in order to determine what type of lag should be applied to the listed price when combining the datasets. In other words, a forecast of when the listed property would be expected to land in the transacted properties dataset would need to be determined. One would expect that the optimal solution lies in spreading the projected sale date over multiple months with a weight applied to that sample based on the predicted likelihood that the transaction settles within that period.

A great deal of research could be explored for this particular topic, however, any promising results yielded by such an investigation would likely be a significantly impactful contribution to the literature. A novel property price index model which is fully automated, is rapidly-computing, operates on frugal data and additionally has the ability to forecast future housing market moves with a high degree of accuracy would likely be considered state-of-the-art, given the rarity of such models available in practice today.

Applying the methodology for the use-case of valuation

A further potential adaptation of the model which could be considered is to employ the same core concept to property valuation. While this is a related field to house price index modelling, they diverge in several ways. Firstly, a house price index has the luxury of averaging across a great deal of samples in the dataset when producing the final result; cancelling out noise and errors. On the other hand, a valuation pertains to one particular property and thus must be considerably more accurate.

The problem of assigning individual valuations to property remains a painfully

manual process today; mortgage lenders frequently send a human surveyor to properties to assess the value which, aside from the resources required, is by no means an exact science. In other cases, lenders typically accept a valuation from the buyer and may cross reference against other recently sold properties in the vicinity to decide whether their valuation is reasonable. Ironically, this is effectively employing the *GeoPrice* index's core methodology to a single case, manually.

It would be interesting to explore whether the *GeoPrice* spatial matching process, potentially in combination with some attribute data, could be used to derive accurate valuations for properties in an automated fashion. Such a system which could produce property valuations automatically within seconds would be of great interest to both lenders and property listing platforms, the former of whom take great risk in conducting their appraisals in such an inexact manner, particularly considering the large size of a typical mortgage.

Another prospective use-case of such a system would be in property taxation. In many nations, property taxes have been proposed as a method to prevent the hoarding of housing. Given that demand far outstrips supply in most nations at present, governments are under pressure to introduce new legislation which disincentivises holding large amounts of property as a speculative asset. At present, this is extremely challenging to do, as the systems to automatically obtain up-to-date valuations on properties are simply non-existent and thus, the administrative burden of attempting to set tax rates is too high a bar to clear.

7.5 Concluding remarks

The examples outlined above are a small selection of the future theoretical research projects which could be undertaken using the *GeoPrice* methodology as a foundation for further investigation. These complement the more practical applications presented earlier in this chapter, which could be launched based on the current state of the model's functionality. Future research will hopefully build on this core facet of spatial auto-correlation further, in order to improve the transparency, timeliness and performance of house price modelling and make these models more accessible and widely available for market stakeholders.

Appendix A

GeoPrice: Building a property price index for the UK market (Further Analysis)

A.1 Price Paid Data: Characteristics

Figure A.1 shows the price distribution of the Price Paid Dataset as a whole, over our sample period, while Figure A.2 demonstrates the difference in distribution of transacted prices for new builds and existing properties.

Figure A.3 demonstrates the distribution of each property type within the regions studied, across our sample period.

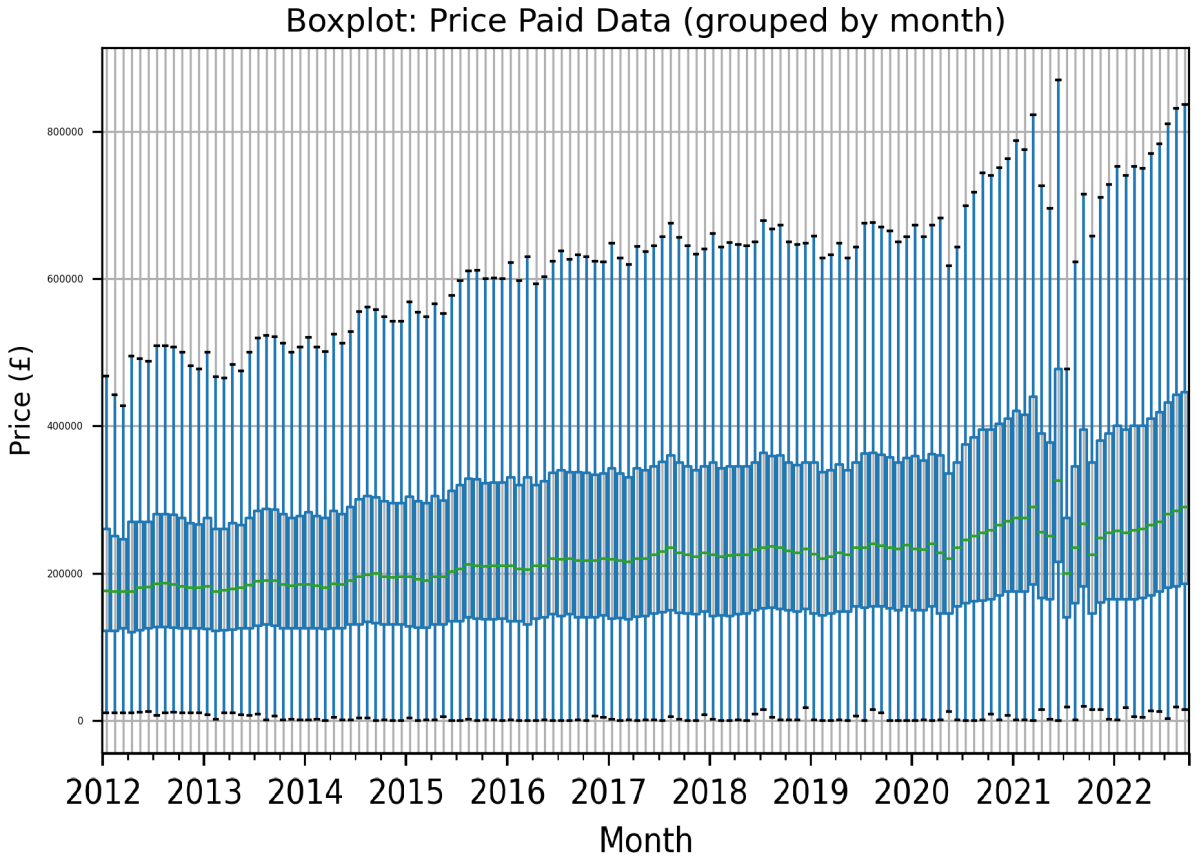


FIGURE A.1: Price Paid Data: Price Distribution from 01-2012 to 09-2022 (inclusive)

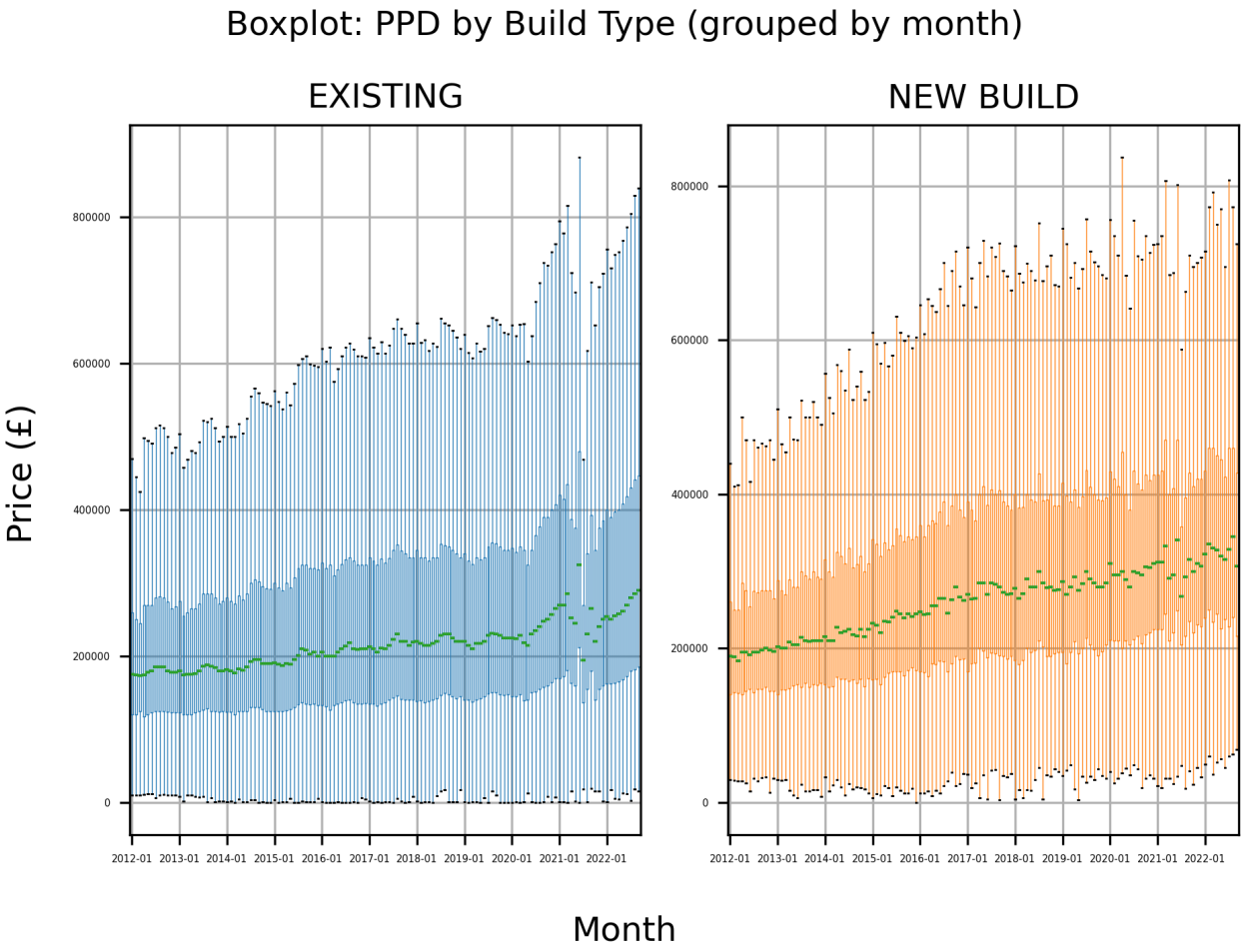


FIGURE A.2: Price Paid Data: Price Distribution from 01-2012 to 09-2022 (inclusive), broken down by build type

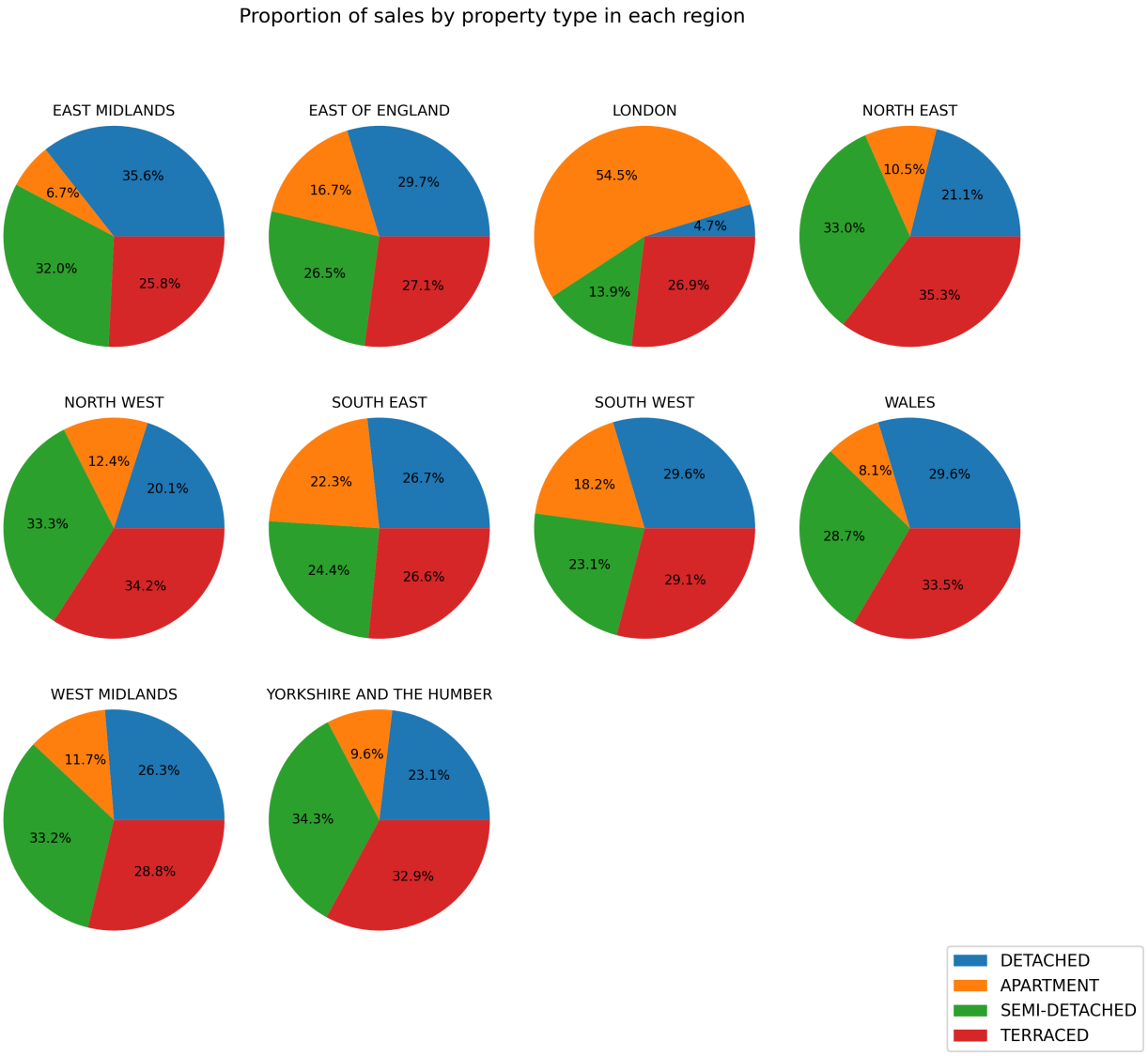


FIGURE A.3: Price Paid Data: Proportion of sales by property type in each region

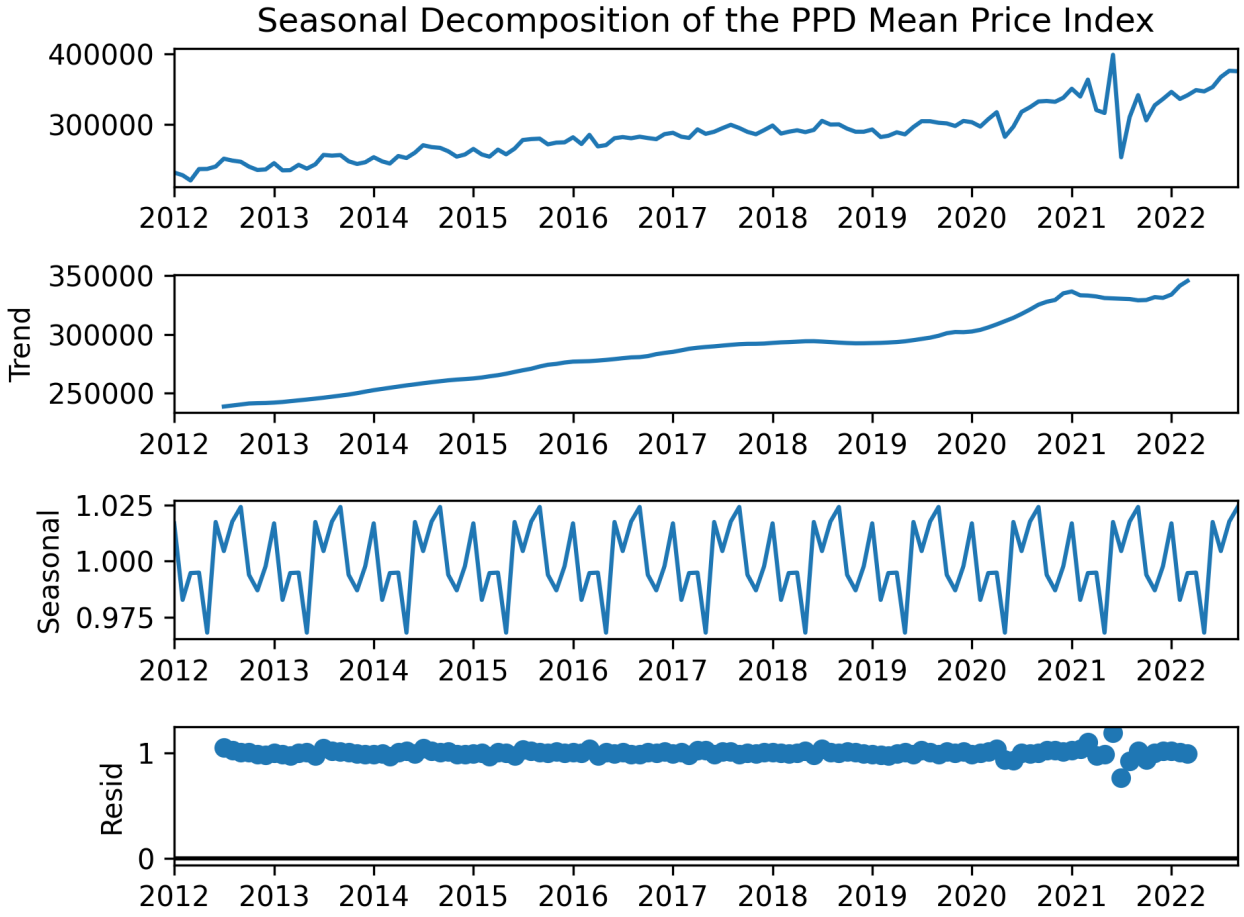


FIGURE A.4: Price Paid Data: Price Seasonality from 01-2012 to 09-2022 (inclusive)

A.2 Price Paid Data: Seasonality

Figure A.4 demonstrates the seasonality of transacted prices in the Price Paid Data. There appears to be a relatively robust seasonal pattern detected, with prices peaking in the late summer and dropping off early in the new year. The residual is highly stable aside from during the extreme volatility period in 2021, indicating a good decomposition of the mean price into trend and seasonal components.

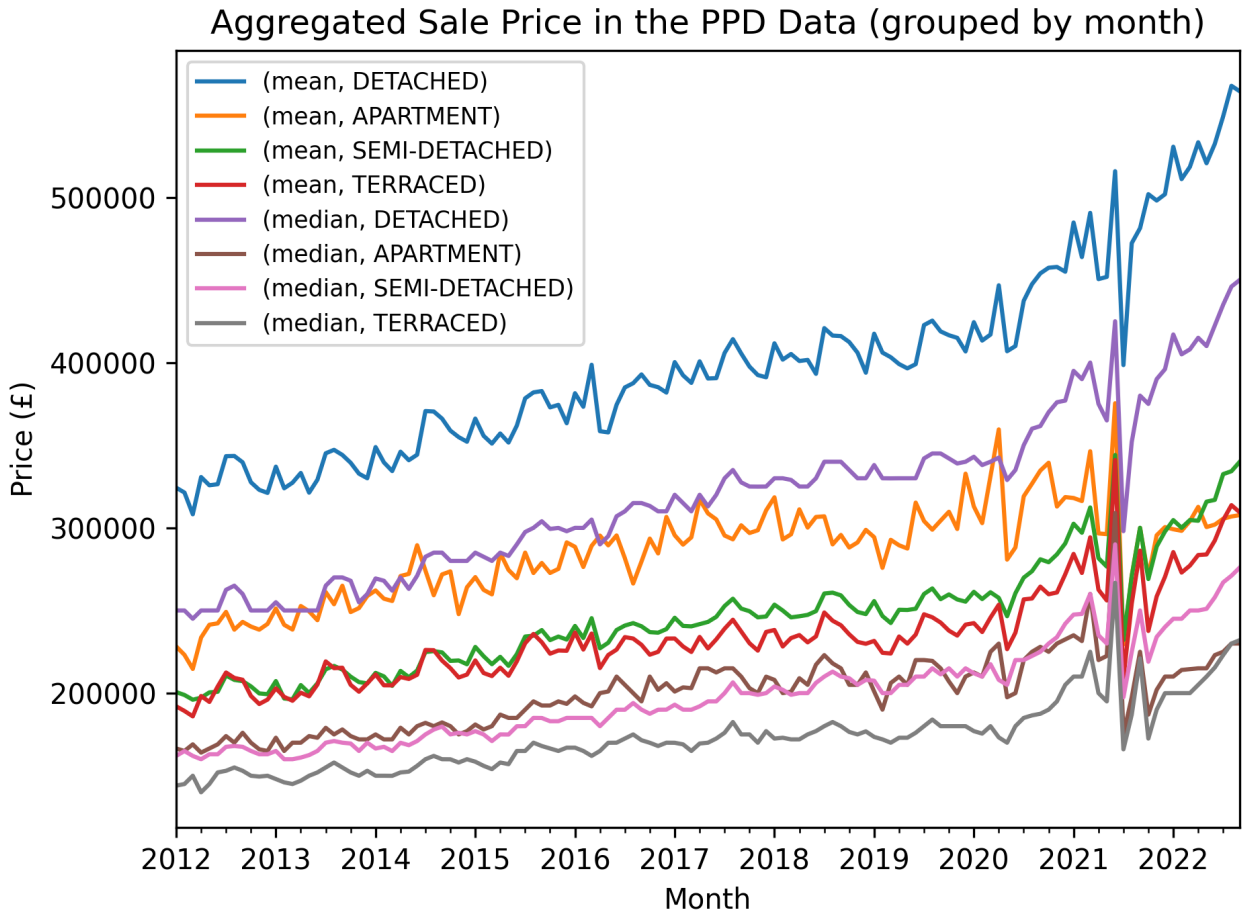


FIGURE A.5: Price Paid Data: Mean/Median Price from 01-2012 to 09-2022 (inclusive), broken down by property type

A.3 Price Paid Data: Mean and Median Indices

Figure A.5 demonstrates the mean and median price index per distinct property type, while Figure A.6 shows a separate index for new builds and existing properties.

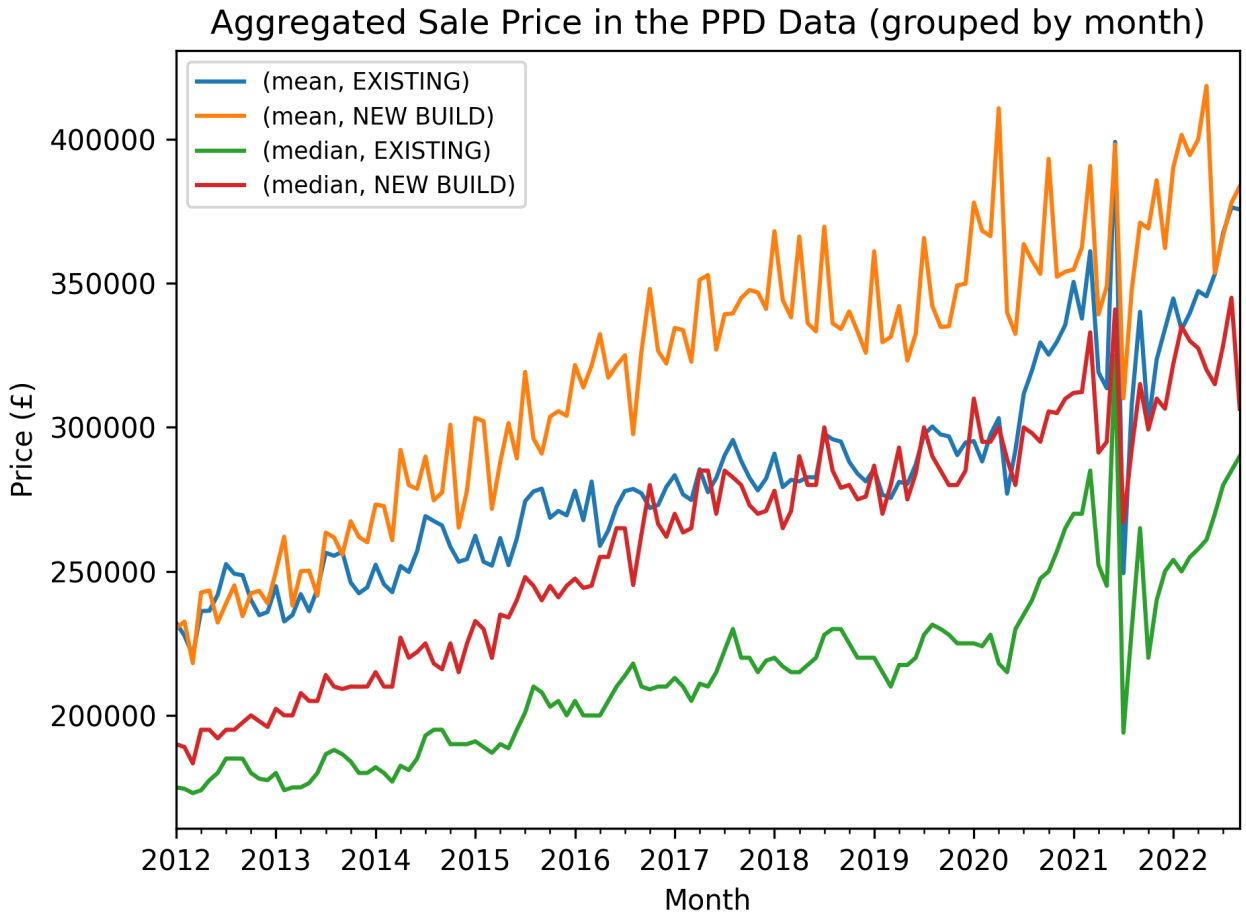


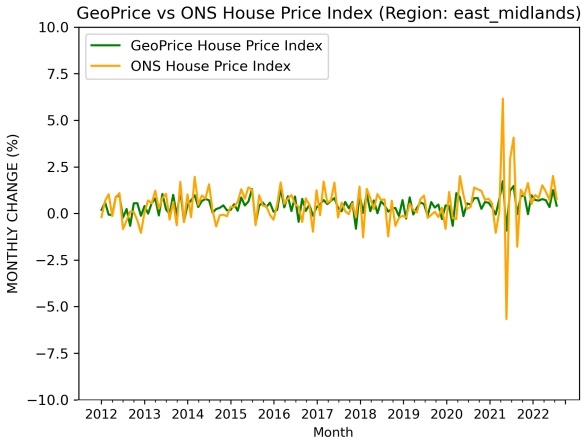
FIGURE A.6: Price Paid Data: Mean/Median Price from 01-2012 to 09-2022 (inclusive), broken down by build type

A.4 GeoPrice with geospatial stratification

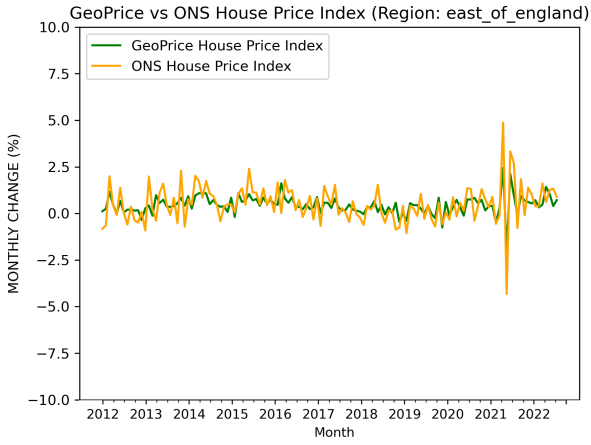
A.4.1 Regional sub-indices

A.4.1.1 Monthly Changes

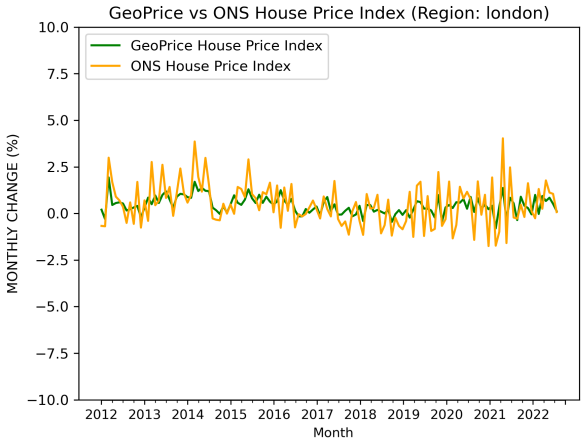
Figure A.7 demonstrates the month on month percentage changes for the GeoPrice index versus the ONS house price index, on a region-by-region basis.



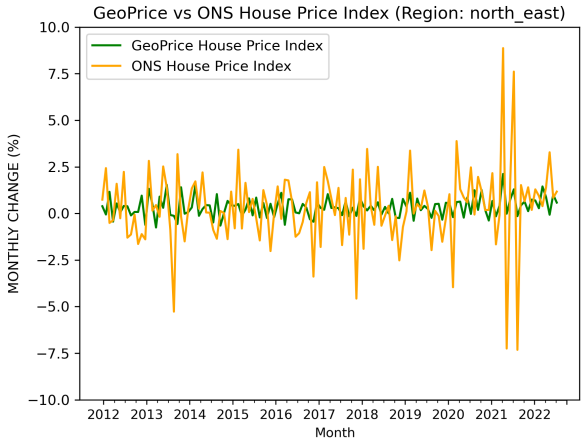
(A) East Midlands Index



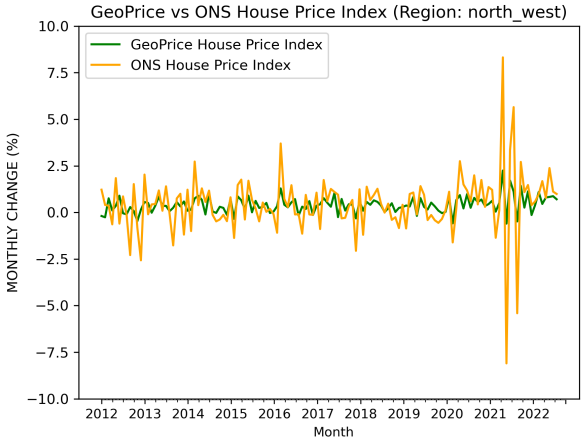
(B) East Of England Index



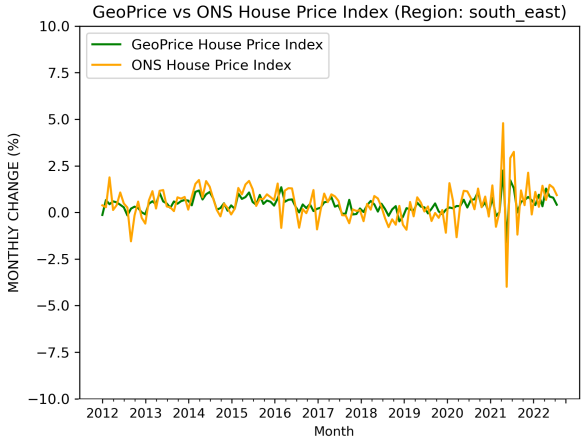
(C) London Index



(D) North East Index

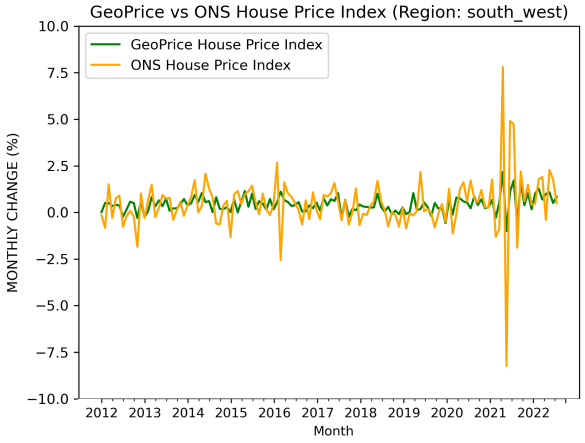


(E) North West Index

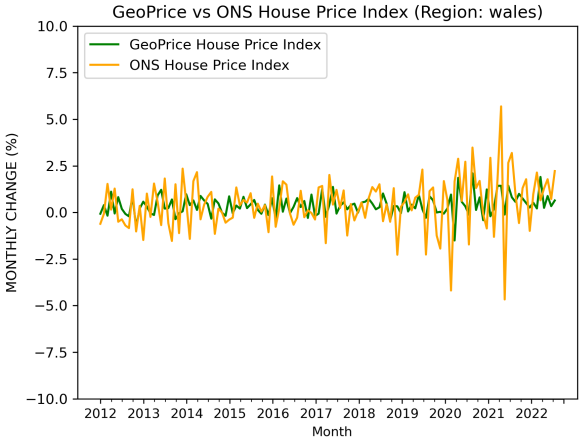


(F) South East Index

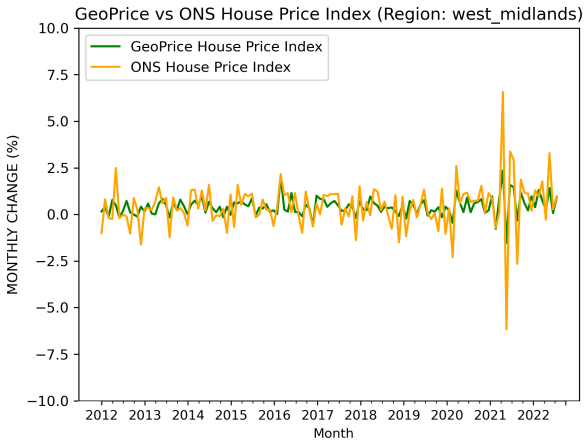
FIGURE A.7: ONS vs GeoPrice House Price Monthly Change from 01-2012 to 09-2022, per region



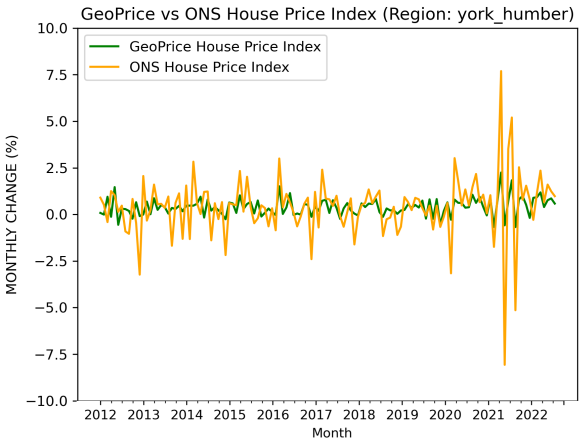
(G) South West Index



(H) Wales Index



(I) West Midlands Index



(J) Yorkshire and the Humber Index

FIGURE A.7: ONS vs GeoPrice House Price Monthly Change from 01-2012 to 09-2022, per region (continued)

A.5 GeoPrice with additional property type stratification

A.5.1 Regional sub-indices

A.5.1.1 Index Levels

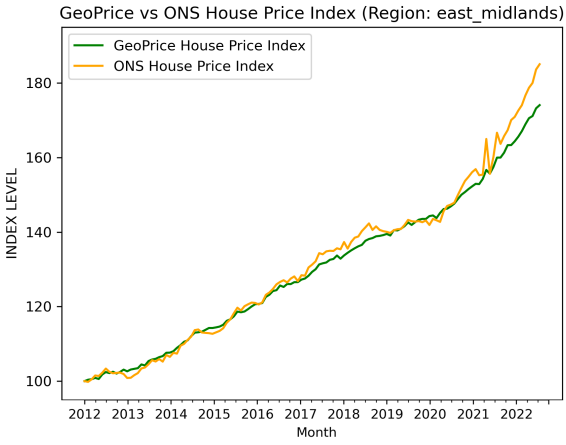
Figure A.8 illustrates the region-by-region breakdown of our GeoPrice house price index versus the ONS hedonic regression index, where the property type has been encoded as an additional stratification attribute in the geohash$^+$ for each sale transaction.

A.5.1.2 Monthly Changes

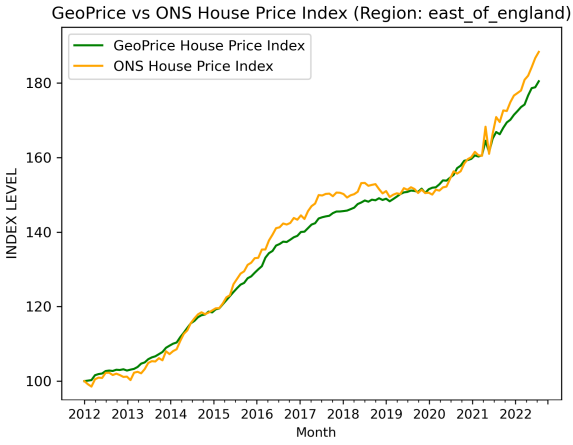
Figure A.9 shows the region-by-region breakdown of the monthly changes of the GeoPrice house price index versus the ONS hedonic regression index, where the property type has been encoded as an additional stratification attribute in the geohash$^+$ for each sale transaction.

A.5.1.3 Smoothness Metrics

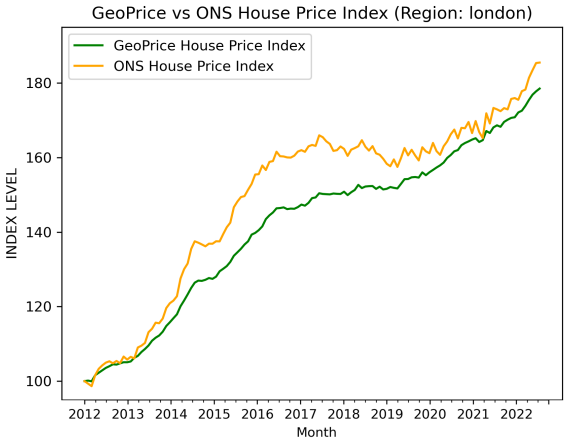
Table A.1 demonstrates the smoothness metrics of the regional sub-indices of both the GeoPrice and ONS house price indices, where the property type has been encoded as an additional stratficiation attribute in the geohash$^+$ for each sale transaction.



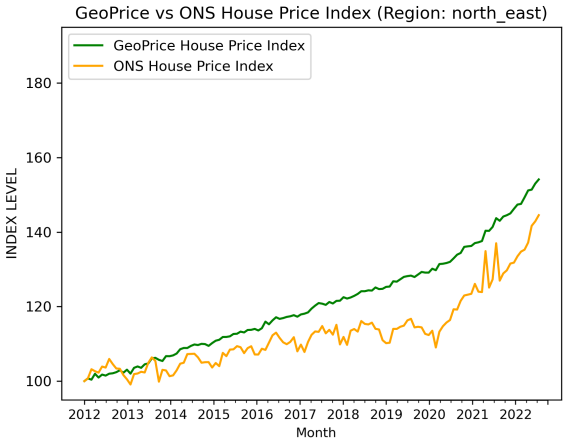
(A) East Midlands Index



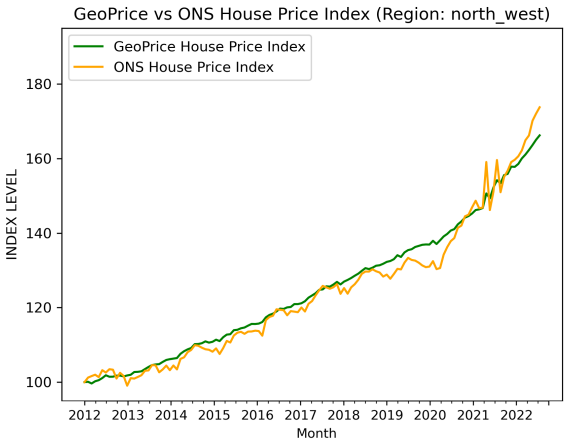
(B) East Of England Index



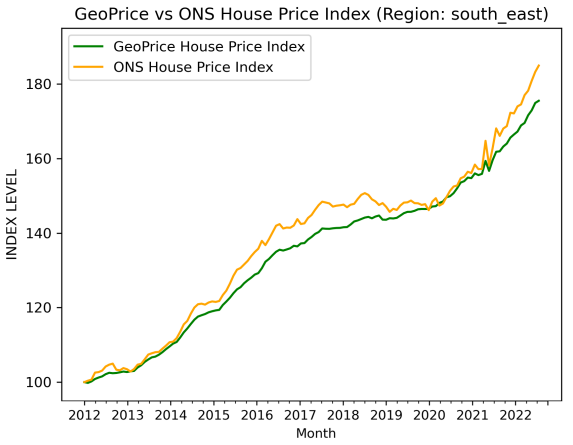
(C) London Index



(D) North East Index

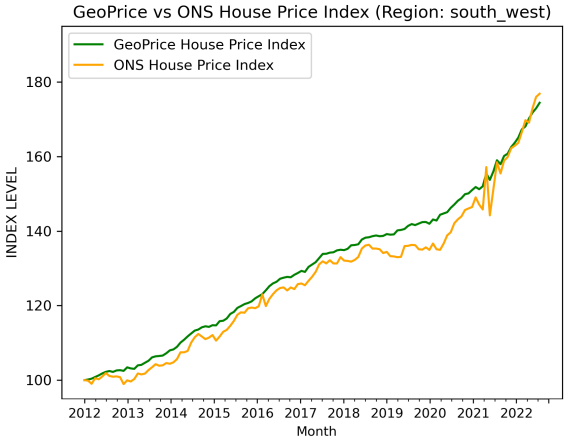


(E) North West Index

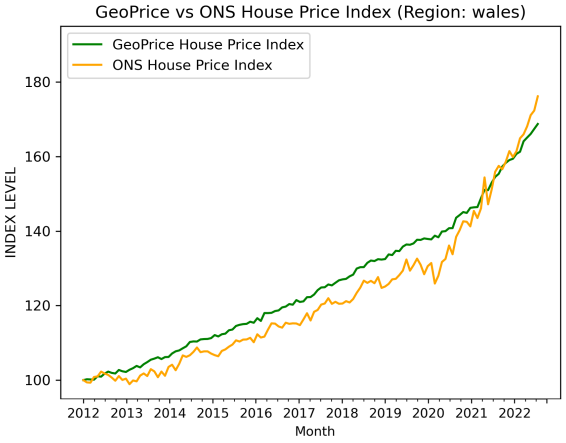


(F) South East Index

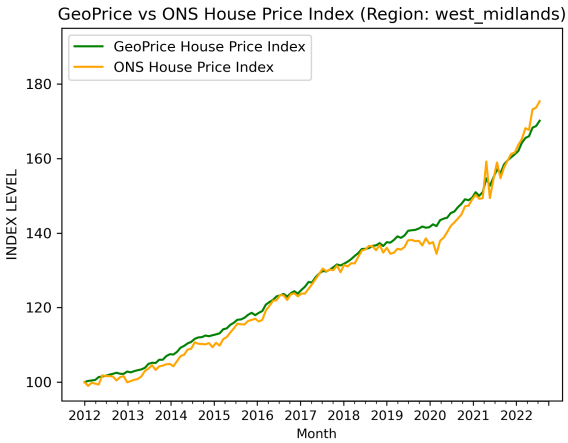
FIGURE A.8: ONS vs GeoPrice House Price Index (w/property type) from 01-2012 to 09-2022, per region



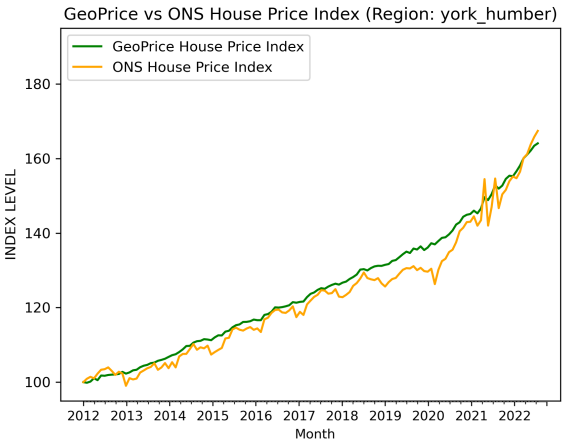
(G) South West Index



(H) Wales Index

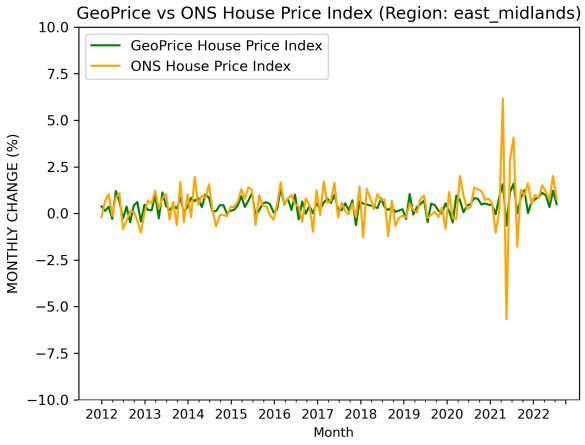


(I) West Midlands Index

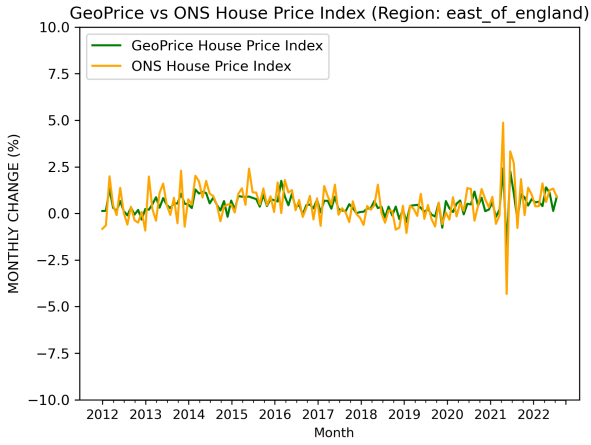


(J) Yorkshire and the Humber Index

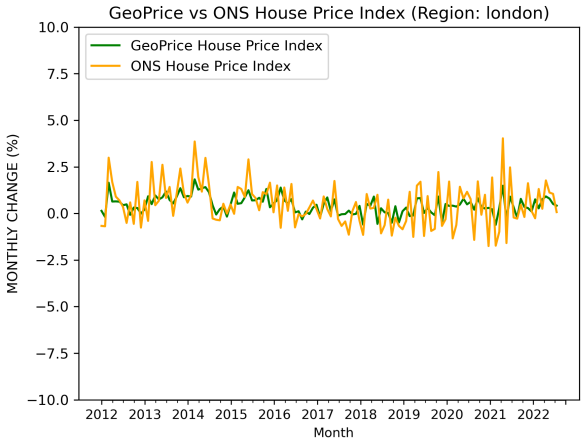
FIGURE A.8: ONS vs GeoPrice House Price Index (w/property type) from 01-2012 to 09-2022, per region (continued)



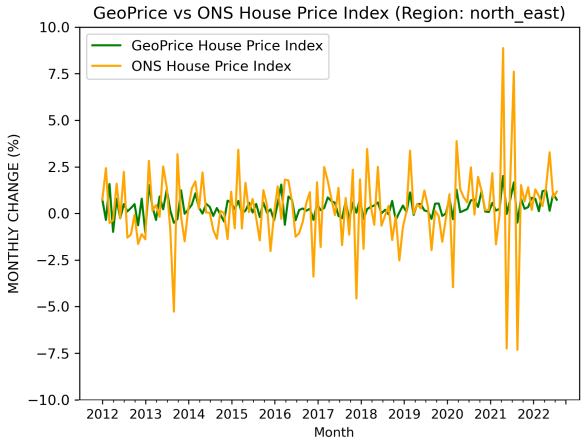
(A) East Midlands Index



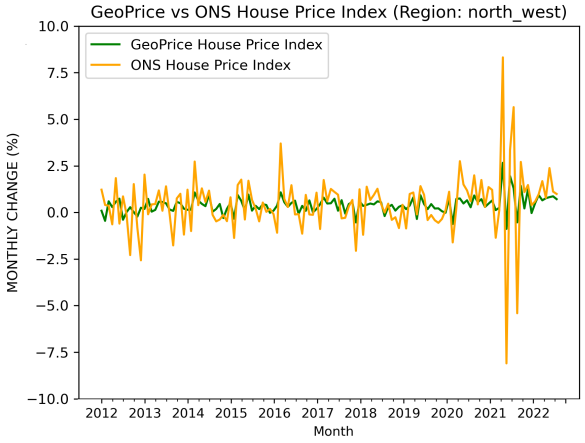
(B) East Of England Index



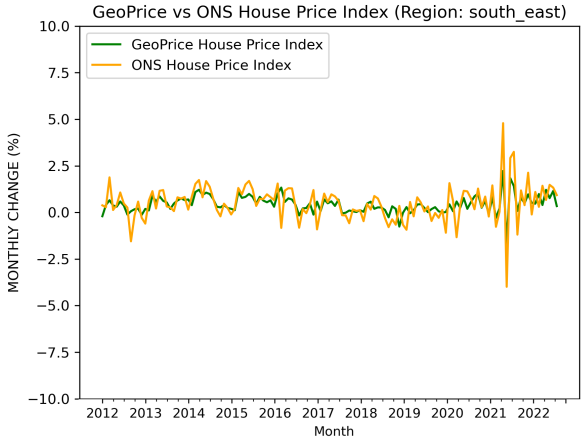
(C) London Index



(D) North East Index

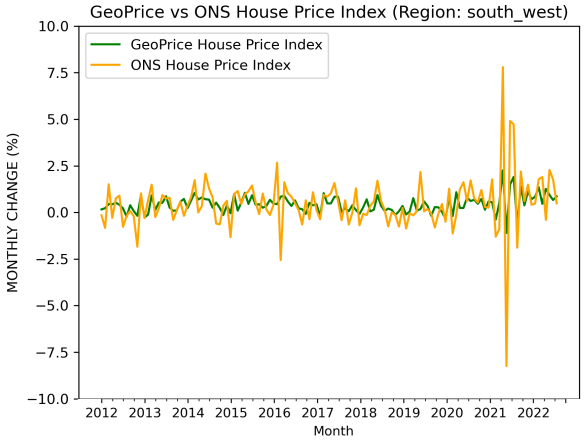


(E) North West Index

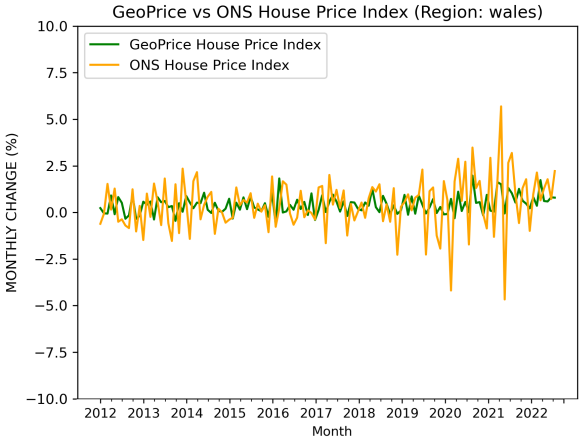


(F) South East Index

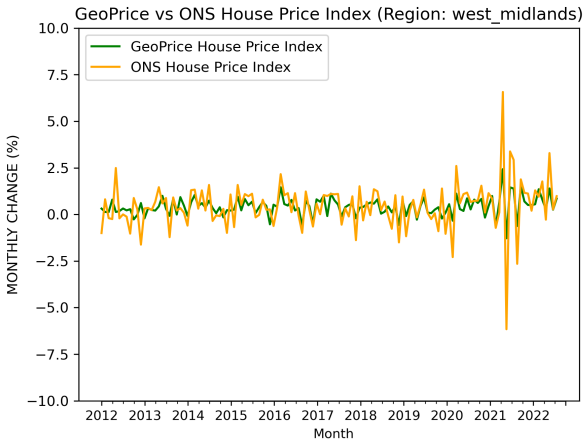
FIGURE A.9: ONS vs GeoPrice House Price Monthly Change (w/property type) from 01-2012 to 09-2022, per region



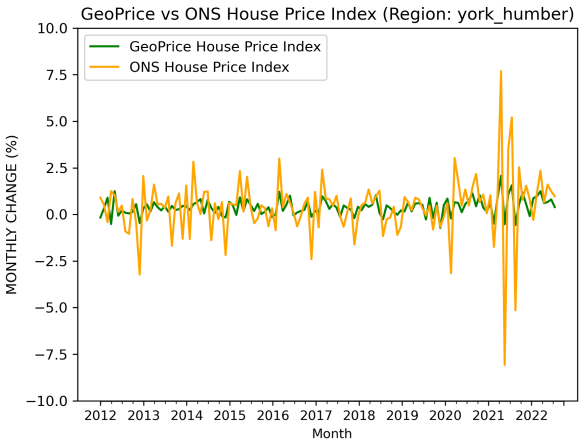
(G) South West Index



(H) Wales Index



(I) West Midlands Index



(J) Yorkshire and the Humber Index

FIGURE A.9: ONS vs GeoPrice House Price Monthly Change (w/property type) from 01-2012 to 09-2022, per region (continued)

TABLE A.1: Smoothness Metrics for ONS and GeoPrice (w/property type) Indices, in each region [UK]

Model	Mean[a]	Median[a]	St. Dev[b]	Min[b]	Max[b]	St. Dev of Diffs	MSM
East Midlands / GeoPrice	0.51	0.47	0.44	-0.65	1.6	**0.68**	**2.09**
East Midlands / ONS HPI	0.85	0.7	1.12	-5.67	6.16	1.84	14.11
East Of England / GeoPrice	0.55	0.46	0.52	-1.86	2.41	**0.78**	**2.66**
East Of England / ONS HPI	0.82	0.59	1.02	-4.32	4.87	1.61	11.25
London / GeoPrice	0.54	0.49	0.49	-0.61	1.83	**0.58**	**1.43**
London / ONS HPI	0.99	0.83	1.16	-1.75	4.03	1.72	12.55
North East / GeoPrice	0.49	0.36	0.54	-1.03	2.01	**0.89**	**3.51**
North East / ONS HPI	1.45	1.17	2.1	-7.32	8.87	3.48	53.76
North West / GeoPrice	0.48	0.45	0.46	-0.88	2.66	**0.74**	**2.44**
North West / ONS HPI	1.12	0.88	1.65	-8.1	8.32	2.77	32.92
South East / GeoPrice	0.5	0.43	0.47	-1.68	2.22	**0.65**	**1.95**
South East / ONS HPI	0.8	0.66	0.97	-3.99	4.79	1.48	9.26
South West / GeoPrice	0.51	0.44	0.47	-1.1	2.26	**0.71**	**2.12**
South West / ONS HPI	0.96	0.65	1.46	-8.24	7.8	2.46	28.79
Wales / GeoPrice	0.49	0.44	0.49	-0.63	1.98	**0.78**	**2.54**
Wales / ONS HPI	1.15	1.01	1.41	-4.67	5.69	2.26	20.7
West Midlands / GeoPrice	0.53	0.47	0.5	-1.28	2.43	**0.8**	**2.72**
West Midlands / ONS HPI	0.94	0.8	1.27	-6.16	6.57	2.15	18.25
Yorkshire H[c]/ GeoPrice	0.47	0.44	0.45	-0.72	2.07	**0.69**	**2.11**
Yorkshire H[c]/ ONS HPI	1.14	0.81	1.64	-8.07	7.69	2.74	33.05

[a] Values quoted are the mean of the absolute monthly percentage changes of each index
[b] Values are reported in percent
[c] *Yorkshire H* refers to the *Yorkshire and the Humber* region

Appendix B

Publication: A robust house price index using sparse and frugal data

Journal of Property Research

Routledge
Taylor & Francis Group

ISSN: 0959-9916 (Print) 1466-4453 (Online) Journal homepage: http://www.tandfonline.com/loi/rjpr20

A robust house price index using sparse and frugal data

Phil Maguire, Robert Miller, Philippe Moser & Rebecca Maguire

To cite this article: Phil Maguire, Robert Miller, Philippe Moser & Rebecca Maguire (2016) A robust house price index using sparse and frugal data, Journal of Property Research, 33:4, 293-308, DOI: 10.1080/09599916.2016.1258718

To link to this article: https://doi.org/10.1080/09599916.2016.1258718

JOURNAL OF PROPERTY RESEARCH, 2016
VOL. 33, NO. 4, 293–308
http://dx.doi.org/10.1080/09599916.2016.1258718

Routledge
Taylor & Francis Group

A robust house price index using sparse and frugal data

Phil Maguire[a], Robert Miller[a], Philippe Moser[a] and Rebecca Maguire[b]

[a]Department of Computer Science, National University of Ireland, Maynooth, Ireland; [b]School of Business, National College of Ireland, Dublin, Ireland

ABSTRACT

In this article, we describe a house price index algorithm which requires only sparse and frugal data, namely house location, date of sale and sale price, as input data. We aim to show that our algorithm is as effective for predicting price changes as more complex models which require detailed or extensive data. Although various methods are employed for determining house price indexes, such as hedonic regression, mix-adjusted median or repeat sales, there is no consensus on how to determine the robustness of an index, and hence no agreement on which method is the best to use. We formalise an objective criterion for what a house price index should achieve, namely consistency between time periods. Using this criterion, we investigate whether it is possible to achieve strong robustness using frugal data covering only 66 months of transactions on the Irish property market. We develop a simple multi-stage algorithm and show that it is more robust than the complex hedonic regression model currently employed by the Irish Central Statistics Office.

ARTICLE HISTORY
Received 28 July 2016
Accepted 6 November 2016

KEYWORDS
House price index; sparse data mining; frugal heuristics; index robustness; central price tendency model

1. Introduction

House price indexes play a critical role in top-level decision-making, and have impacts on investment decisions by both the private and public sectors (Plakandaras, Gupta, Gogas, & Papadimitriou, 2015). House owners, bankers and policy-makers all pay close attention to relative price levels and the magnitude and direction of price changes in both regional and local markets (Costello & Watkins, 2002; Leishman, 2009; Munro, 1987). This information can be useful in forecasting inflation, economic output and real GDP growth (Case, Quigley, & Shiller, 2005; Forni, Hallin, Lippi, & Reichlin, 2003; Gupta & Hartley, 2013; Gupta & Kabundi, 2010; Stock & Watson, 2004)

House price indexes are also important for academic research aimed at understanding the dynamics of the market, and for investigating issues of societal relevance, such as housing affordability and price bubbles (Bourassa, Hoesli, & Sun, 2006). The study of index robustness is particularly relevant in the contemporary financial environment, given the recent price volatility in international housing markets and the prominence of housing market debt instruments as a primary cause of the global financial crisis (Goh, Costello, & Schwann, 2012). The possibility of hedging against housing risk (e.g. Englund, Hwang, & Quigley, 2002; Shiller, 2003) depends on access to extremely accurate price indexes.

CONTACT Phil Maguire pmaguire@cs.nuim.ie

Most house prices indexes require either extensive data which stretch back decades or else detailed data, which describe numerous features of each home. Our aim in this article is to develop an algorithm which requires only a few months of transactions (sparse data) and the barest of details (frugal data). We hope to show that such algorithms can match the robustness of more complex data-intensive methods. If feasible, such techniques would have numerous advantages over the systems currently in use. For a start, they would be less labour-intensive, relying on information scraped automatically from webpages, with no need for the input of expert statisticians. Second, they would be more responsive, giving consistent up-to-date information about house price changes. At the moment, statistics offices, such as the Irish Central Statistics Office (CSO), release information only once a month, with nearly a month of delay. An automatic algorithm could recompute the changes every few minutes, using not only sale prices, but also asking prices gleaned from online property websites, thus capturing immediate shifts in market sentiment.

2. House price index approaches

We begin by providing an overview of existing strategies for determining house price indexes. Given the importance of house price movements and the voluminous associated literature (see Hansen, 2009, for an overview), it is perhaps surprising that no consensus exists on how an index should be constructed. When comparing house price index models, researchers are faced with numerous data and methodological issues which stand in the way of constructing an accurate index (Goh et al., 2012). First of all, housing markets are highly illiquid. Due to substantial search, transaction and relocation costs, only a fraction of the total housing stock is sold each year. The 2008 financial crisis and subsequent Irish property crash led to a much lower number of transactions than usual for this period (see Lyons, 2015). For example, according to the stamp duty returns maintained by the Irish Property Services Regulatory Authority (PSRA), in the period January 2010–July 2015 less than 150,000 properties were transacted, out of a total of 1.65 million (9.1%), implying that the average house is transacted once every 60 years.

Another problem is that the properties being sold have varying characteristics which are affected by geographical and temporal factors, introducing potential bias into the sample selection. Houses are also subject to quality change over time, which can also vary by area.

To overcome the problem of small sample size, data are often pooled arbitrarily into broad representations of time and geography. The assumption here is that the pooled sample will produce price indexes that are statistically equivalent to those that would have been obtained from the smaller constituent subsamples. This must be done carefully, as excessive pooling of data for house price index construction can lead to biased price index estimates (Englund et al., 2002; Goh et al., 2012). Developing and maintaining an unbiased index according to best international practice is a complex and demanding process (see de Haan & Diewert, 2011). In the following sections, we describe the three main techniques used for deriving house price indices, namely hedonic regression, repeated sales and adjusted-mix median.

2.1. Hedonic regression

The hedonic modelling method is used to construct house price indexes in Ireland and in the UK. The central idea, originally introduced by Kain and Quigley (1970), is that of determining the quality of a given house by decomposing it into its constituent characteristics, then estimating the contributory value of each characteristic (e.g. number of bedrooms, distance to city centre, plot size, etc.). The results of the regression indicate the changes in property values for a unit change in each characteristic, assuming that all the other characteristics are held constant.

The advantage of the hedonic approach is that physical attributes such as location, age and size are introduced into the regression model, and their net contribution to the market price is estimated (Bourassa, Hoesli, & Sun, 2006). Although hedonic regression is found in the literature to provide a good fit with the data (e.g. Goh et al., 2012; Shimizu, Nishimura, & Watanabe, 2010; Wallace & Meese, 1997), the disadvantage is that it requires a lot of data, which are not always available, or can be impractical to obtain. Many of the attributes that can be expected to influence the price of a property, particularly neighbourhood and location variables, are often not available, and other relevant attributes may go undetected (Case, Pollakowski, & Wachter, 1991). Hedonic models are relatively complex to interpret, and require a high level of statistical knowledge and expertise (Bourassa et al., 2006). The fact that there are many free parameters available to be tuned also increases the risk of overfitting (see Heene, Coyne, Francis, Maguire, & Maguire, 2014).

2.2. Repeat-sales

The repeat-sales method is another popular house price index technique that controls for the heterogeneity of properties. The method, originally developed by Bailey, Muth, and Nourse (1963), and further enhanced by Case & Shiller (1987), holds house quality constant by measuring the same asset in two different periods. As a result, there is no need to include the property attributes in the model; transaction prices and property address are sufficient. This index methodology has evolved into the most widely used and reported US house price index.

One drawback of this approach is that, because repeat-sales models consider only dwellings with multiple transactions, they require large amounts of data stretching back in time (de Vries, de Haan, van der Wal, & Marin, 2009). Only a fraction of transactions at any given time period will have matching historical sales, and this sample may not be representative of the market as a whole, leading to aggregation biases (Dombrow, Knight, & Sirmans, 1997).

For example, frequently transacted houses may have some idiosyncratic characteristics that make the owner eager to sell (Sommervoll, 2006). In contrast, frequent transactions might equally indicate that a property has some characteristics that make it easy to resell. Further complicating matters, an analysis carried by Case, Pollakowski, and Wachter (1997) suggests that frequently resold houses tend to appreciate more than those that are less transacted. Short holding periods may indicate significant renovation activity has occurred between sales, therefore violating the assumption of constant quality. Costello (2000) demonstrates that the accuracy of repeat-sales indexes improves significantly when long holding periods (more than one year) are used in estimation of repeat-sales indexes.

There are a number of other weaknesses associated with repeat-sales indexes. One of the most serious is revision, which means that past values of the index are perturbed and revised by present-day information (Baroni, Barthélémy, & Mokrane, 2005). Additional sales reverberate on the index values because new sales pairings provide information on price movements which go beyond the information originally available.

2.3. Central-price tendency methods

The idea of central-price tendency models is that, by aggregating large amounts of data, random noise will naturally tend to cancel out following the law of large numbers, leaving a reliable signal. This approach is far less data-intensive than either hedonic regression or repeat-sales, requiring neither detailed information about properties, nor extensive historical data-sets. One feature that central-price tendency methods do assume is that the data being aggregated are drawn from the same distribution, and cannot be subdivided into different distributions which might be differentially affected over time.

In the US, the index published by the National Association of Realtors is based on median prices (Bourassa et al., 2006). Although such indexes are simple to construct, there is little control for robustness (Case & Shiller, 1987).

Central price tendency models are often criticised as they do not control for the attributes of houses sold either directly in estimation, or indirectly by sample selection (Goh et al., 2012). This can result in inaccurate indexes, susceptible to variations in the mix of houses sold from period to period in a particular region.

Richards and Prasad (2008) argue that stratifying the full sample by suburb, and then taking the simple average of the median sale prices across each suburb, yields price index estimates that are not significantly different from hedonic regression. Given the effectiveness of this strategy for stratification, Richards and Prasad (2008) suggest that the marginal benefits of the more complex and data-intensive methods, such as hedonic regression and repeat-sales, are not justified.

2.4. Comparison of approaches

Goh et al. (2012) directly compared these three different strategies and concluded that hedonic regression models give the best performance. Two variants of the hedonic model were used, namely the standard explicitly intertemporal model and the 'imputed' cross-sectional model. The latter was found to outperform all other index models, matching the findings of previous studies (e.g. Diewert & Hendriks, 2011). Schwann (1998) observed that price indexes constructed using standard hedonic regression are the most robust to finer levels of temporal and geographic disaggregation. He also proposed a time series model employing a stochastic structure for hedonic parameter evolution, which achieves further stabilisation in sparse markets.

The mix-adjusted median was the next best performer in Goh et al.'s (2012) study, with repeat-sales faring the worst, which, given its prominence in the evaluation of the US housing market, is surprising. Shimizu et al. (2010) found that the repeat-sales approach measures market turning points later than the hedonic approach, the former being more than two years delayed in the case of the Tokyo housing market. Wallace and Meese (1997)

also concluded that repeat-sales and other hybrid methods produce less reliable estimates of price movements than the hedonic approach.

Goh et al.'s (2012) results reject the null hypothesis of equality between mean hedonic characteristics for the samples of single-sale and repeat-sale dwellings, revealing that repeat-sales are not representative of the market in general. Houses sold more than once are significantly smaller, have fewer amenities and are of poorer quality, supporting the observation that repeat-sale dwellings are generally sold at a discount to non-repeat-sales.

Goh et al. (2012) reported that, although the performance of the mix-adjusted median was merely 'modest', the method deserves some credit because of its simplicity and transparency. Because it assumes that all houses in a given location stratum are drawn from the same distribution of hedonic quantities, there is no need to identify hedonic attributes of individual houses, collect large amounts of data or carry out any esoteric statistical procedures. Goh et al.'s (2012) findings support Richards and Prasad's (2008) claim that, in absence of information on hedonic attributes, the mix-adjusted median is likely to be the best alternative. Our aim is to investigate whether the central tendency approach can be enhanced to the point where it can compete with, or even outperform, hedonic regression, as applied to the Irish housing market.

3. Case study: the Irish residential property price index

The Irish property market is an example of a relatively sparse data-set. For the period 2010–2015, there were only, on average, 2,200 transactions per month nationwide, motivating the development of techniques for achieving high levels of robustness from small amounts of data.

Currently, property price changes in Ireland are reported only monthly, more than three weeks into the new month, and only broken down for two subregions, Dublin, and outside Dublin, for apartments and for all properties. The Residential Property Price Index (RPPI) is compiled by the CSO, using a hedonic regression 12-month rolling time dummy model (O'Hanlon, 2011). In addition, the monthly results that are released to the public are based on a rolling average of the previous three months, thus enhancing the smoothness of the time series. However, the disadvantage of such artificial smoothing is that the RPPI loses responsiveness to changing market conditions, and can appear misleadingly precise to observers who are not aware of the use of rolling average.

Currently, there are two significant sources of data available for compiling a house price index in Ireland. The first is mortgage returns, which are filed by all lending agencies for properties whose purchase was partly funded by a mortgage. Irish mortgage lenders are required, under Section 13 of the Housing Act 2002, to submit monthly mortgage returns to the Department of Environment, Heritage and Local Government containing data on both mortgage approvals (occurring where a formal letter of mortgage offer has issued) and mortgage drawdowns (O'Hanlon, 2011).

The advantage of this information source is that it carries detailed information about the property, such as the number of bedrooms, the floor area, year of build, plot size, etc. The disadvantage is that not all properties are purchased with a mortgage, hence the sample is unrepresentative. As property prices rise, more people are in a position to trade down to cheaper properties without a mortgage. In addition, lending restrictions following the property crash have led to an increase in cash transactions: from 2010 to 2014, mortgages

on house purchases fell from 88% to only 50% (Dalton & Moore, 2014). Furthermore, 68% of mortgage returns contain errors, such as a missing year of construction, missing number of rooms or missing plot size. Missing, erroneous and implausible values are imputed by the CSO (O'Hanlon, 2011).

The fact that half of transactions are missing from the mortgage records is not necessarily a problem. If 50% of data is randomly removed from a data-set, it has at most a mild effect on the robustness of any index computed from it, amplifying the standard error by $\sqrt{2}$. What matters more is when the missing data are not a random sample, but have some relationship with the rest of the data, which is not taken into account by the model.

For the mortgage data, it is likely that the missing 50% is not a representative sample. Cheaper investment properties are more likely to be transacted in cash, without a mortgage. By contrast, the purchase of larger family homes is more likely to require a mortgage. For this reason, even if the CSO's hedonic regression achieves high goodness-of-fit statistics, the performance is potentially taking place within a biased sample, meaning that goodness-of-fit is not a reliable measure of robustness.

A second source of available information is stamp duty returns, maintained by the PSRA. This publicly available online database reports the date of sale, sale price and address of every property sold in Ireland since 1 January 2010, with a typical latency of around 10 days. The disadvantage of this information source is that it includes no information on the property. Even the addresses can be unreliable, as Ireland has only recently introduced a postal code system, which is yet to be adopted by the PSRA. Although the large majority of returns are lodged immediately, some are delayed by up to 3 months before being submitted to the National Stamp Duty Office. A final disadvantage is that as well as including market sale transactions, the records also include a small proportion of non-sale transactions (e.g. properties that are inherited), which could potentially bias a house price index because the values involved are much lower.

In the case of stamp duty returns which are delayed, it seems reasonable that the subset of records which get delayed is a random selection: the type and location of property purchased should have no predictable relationship with the issue of whether the associated stamp duty is lodged promptly or not. As regards the non-market transactions, if these occur randomly through time periods and geographic locations, then this noise should tend to cancel out for large data-sets using a central-tendency approach.

According to O'Hanlon (2011), the failure of stamp duty returns to collect details on the characteristics of properties rules out the possibility of carrying out an appropriate level of mix-adjustment. He concludes that the Property Price Register can only be of benefit to users with detailed knowledge of the characteristics of specific properties (such as local inhabitants, local estate agents).

In this article, we investigate the hypothesis that a frugal data-set recording only address, date of sale and sale price is sufficient for deriving an index of equivalent robustness to the RPPI currently produced by the CSO. Addresses can, with high reliability, be converted to GPS locations through freely available mapping systems, such as Google Maps. This geographic positioning should permit mix-adjustment and stratification using appropriate central-tendency strategies.

Heene et al. (2014) have argued that simple models with fewer parameters are better suited to modelling complex phenomena, because they minimise the risk of arbitrary overfitting. If an automated frugal data model can match the performance of the CSO's

hedonic regression it would have many advantages, requiring no labour or expense to maintain, and being available with 10 days latency, rather than 3 or 4 weeks.

But first we need to set the rules by which the competition will be decided: we must define index robustness.

4. Measuring robustness

Despite being of critical importance for research in this area, the issue of robustness has received little attention (Goh et al., 2012).

What does a good index look like? According to Chandler and Disney (2014), it is surprisingly hard to identify what exactly house price indexes are intended to measure. Even the language used by the organisations compiling the indexes is vague. For example, the UK's Office for National Statistics (ONS) states that 'the aim of the ONS House Price Index is to measure the change in the average house price for owner-occupied properties in the UK'. But what does 'average' mean? This ambiguity creates difficulties in assessing the relative accuracy and robustness of different index models.

The 'true' house price trend is unobservable, since identifying 'true' house prices would require measurement of the total stock of housing in the local market (Goh et al., 2012). Wallace and Meese (1997) addressed the problem by assuming that the 'true' index can be proxied by the median house price, though Goh et al. (2012) argue that this is contrary to a large body of literature which argues against the application of the median (e.g. Case & Shiller, 1987; Hansen, 2009).

Case and Szymanoski (1995) and Richards and Prasad (2008) developed methods for comparing various models by directly comparing goodness-of-fit statistics. However, Sommervoll (2006) argues that, due to the risk of overfitting, goodness-of-fit statistics can be misleading, especially where indexes are estimated at high levels of disaggregation or for sparse data. Serious mis-measurements may occur, even in cases where the statistical diagnostic tools like R^2, t-values and standard deviations indicate good explanatory power.

The underlying problem with goodness-of-fit is that it fails to account for complexity: models should somehow be penalised for the number of degrees of freedom they exploit to achieve a certain level of fit. In the light of this, model performance is better evaluated through *forecast error*. One way to test forecast error is to randomly divide a data-set of property transactions into two halves: if the index is robust, both halves should yield the same index value.

Following this idea, Goh et al. (2012) adopt a within-sample cross-validation strategy. They randomly select a 75% subset of transactions and evaluate how well the index computed on this selection predicts the sale price of properties in the other 25% subset. The closer the match, the more robust the index.

A problem with Goh et al.'s (2012) test for robustness is that a single iteration of cross-validation is not reliable. For example, two random halves might by chance produce close agreement, where nearly every other partition would have resulted in diverging values. Specifically, the values returned from a single implementation of the cross-validation technique are themselves drawn from a distribution, with an associated mean and standard deviation. The process must be repeated many times to identify a reliable sampled mean. As the number of partitions n increases, the reliability of the robustness value increases with order $\sqrt{n}$.

While Goh et al.'s (2012) test can provide a weak heuristic for assessing robustness, it cannot provide the basis for a definition, since it is easy to construct an index which is not robust, yet does well at the test. For example, we could hardcode an algorithm that outputs 100 if the number of transactions in the sample is even, and 99.999 if the number is odd. The agreement will always be very high for any random split, and this agreement can be boosted to any arbitrary level by adjusting the hardcoded values. And yet the index is uninformative.

We propose a minor refinement of Goh et al.'s (2012) test, which can serve as a definition for robustness. Given two competing indexes, the more robust index is the one which, when run repeatedly on two random partitions of a given data-set, produces a pair of values which, on average, are closer to each other than those produced by the other index. We also restrict the set of functions to those which vary monotonically with the change in any sale price in the set (i.e. if any sale price is altered, the function output must either stay the same or move in the same direction). To formalise this mathematically, a valid index function is a computable function $i : R^n \rightarrow R$ that is monotonic, i.e. for all $\epsilon > 0$ and for all $x \in R^n$, it holds that $i(x_1 + \epsilon, x_2 + \epsilon, \ldots, x_n + \epsilon) > i(x_1, \ldots, x_n)$. This is close to Goh et al.'s idea of cross-validation prediction, except that it knocks out the pathological examples, as highlighted above, where an index ignores the input, and always produces the same hardcoded output.

In practice, the most robust index is the smoothest index. Our argument is as follows: given an index, some component of the monthly price fluctuation is due to random sampling error, and the remaining component is due to genuine shifts in market sentiment. We want to eliminate as much of the background noise as possible, thus allowing us to tune in to the signal of the market itself. Comparing discrepancies between successive months is similar to Goh et al.'s (2012) idea of comparing different samples drawn from the same month: the goal is to develop an index with the smallest discrepancies.

Changes in market sentiment have a lower frequency than that of background noise: for example, we expect the market to move in cycles, with prices drifting consistently upwards for months, then drifting consistently downwards during a recession (see Agnello & Schuknecht, 2011). In contrast, sampling error stemming from the construction of the index will jump randomly from month to month. While changes in house prices have momentum, sampling error does not (thus explaining why the CSO chooses to publish three-month rolling averages). Because of this differential in frequencies, smoothness acts as an indicator of noise filtering. The smoother the trending of the index (i.e. the greater the extent to which changes in successive months agree with each other), the smaller the noise component, and the higher the reliability of the remaining signal. Accordingly, we will evaluate index robustness in terms of the average absolute monthly change in market momentum; a steadily rising index would have an average change of zero.

According to Wang and Zorn (1997), an index should be defined by its use in practice, rather than by the more complex, higher level concerns of statistics and models. They find that much of the debate over index methodology can be distilled to implicit and largely unrecognised disagreement as to the intended application.

Taking Wang and Zorn's recommendation into account, we can express the above mathematical definition in terms of a clear practical application: the most robust index is the portfolio that investors would naturally seek to hold if house price indexes were openly traded in a prediction market (as recommended by Englund et al., 2002 and Shiller, 2003).

Investors seek to hold a portfolio which is as diversified as possible, thus minimising risk, while maintaining return (see Maguire et al., 2014). For example, a diversified index, such as the S&P 500 index for the US stock market, should have a better risk to reward profile than any of its constituents, or indeed, any subset of its constituents (c.f. Maguire et al., in press). This is why investors seek to hold the S&P 500 index, and why it provides the gold standard for financial models.

Choueifaty, Froidure, and Reynier (2013) propose that portfolio diversification is related to volatility, and can be evaluated by the extent to which independent sources of information combine to smooth the overall volatility of a portfolio. For example, if numerous house price indexes published by different organisations were freely available to trade, investors would naturally hold the portfolio which minimises overall volatility, thus in effect creating a more diversified super-index with a better risk to reward profile than any of its individual constituents. In sum, the optimal house price index, the one that would be most traded and hence most quoted in the media, is the smoothest house price index.

In the following section, we describe an algorithm developed to meet this objective standard for index robustness, which functions on sparse (no long-term historical records) and frugal data (only location, date and price).

5. Algorithm

Our algorithm involves several stages of processing the online data provided by the PRSA. First, we collected all the available data, stretching back from January 2010 to the end of June 2015. Google Maps API provided the best option for geocoding the addresses into GPS co-ordinates. The service has a rate limit of 2,500 requests per day, so the process was carried out automatically over a period of 2 months.

Approximately 90% of addresses were successfully converted, giving us the GPS co-ordinates, date of sale and sale prices for 147,635 unique transactions. These transactions were analysed in monthly sets, with January 2010 providing the base index of 100. There were considerable differences in the number of transactions per month, from a low of 677 in January 2011, to a high of 3,894 in December 2014.

The first index we calculated was based simply on the raw average property price for each month. The time series of price changes for this index had an average monthly change in momentum of 12.40%. Subsequently, we computed the raw monthly median. The monthly shift in momentum of this index was lower than that of the raw average, at 8.42%.

5.1. Stage 1: Filtering

The stamp duty return data show that whole housing estates and blocks of apartments are sold in bulk at the same time, greatly distorting the average price in a given month. The next stage of the algorithm was to remove these distortions.

Making use of the geographic co-ordinates, we eliminated any property transaction for which there was another transaction within 100 m in the same period of 48 h. This eliminated all bulk sales, with the number of valid transactions being reduced by 14.4% to 126,444. The average monthly change in momentum of the median of this filtered subset of transactions was lower again, at 6.37%.

5.2. Stage 2: Mix-adjustment through proximity voting

A potential problem with median-based approaches is that fluctuations in the relative number of properties sold in unrepresentative regions can have a dramatic effect on the median, even when there is no increase in price. For example, if twice as many properties in Dublin are sold as usual, a region in which the value of most properties is higher than the national median, these sales will act to drag the median upwards, despite no actual change in price.

Richards and Prasad (2008), for example, found that, based on a database of 3.5 million transactions in the six largest Australian cities, compositional shifts between higher and lower priced parts of cities led to much volatility in unadjusted median prices. Similarly, realtors in the US report that median house prices rise in the summer: most families with children, who typically buy more expensive homes, time their purchase based on school year considerations (Richards & Prasad, 2008).

To enhance robustness, it is important to control for the mix of properties which are sold in any particular month. What we want is to identify a subset of the given sample which is more representative of the houses in the market as a whole. Specifically, we want the analysis set to be as spatially autocorrelated as possible with the set of historical records, featuring the same relative distributions of transactions in different regions of the country, and the same types of properties within those regions. Spatial autocorrelation arises in housing data due to the proximity of units that are the same or among contiguous units (Hamid, 2001). In general, properties in close proximity tend to have similar structural characteristics, such as size of living area, dwelling age and design features (Ismail, 2006). The similar quality of proximate properties is a natural consequence of the fact that they tend to be developed at the same time (Gillen,Thibodeau, & Wachter, 2001). Residents in the same neighbourhood may also follow similar commuting patterns, and share the same neighbourhood amenities such as public schools and shopping centres (Ismail, 2006).

In light of this, we developed a system for enhancing autocorrelation based on geographical proximity to a historical target set of transactions. Specifically, we eliminate the 10% least representative properties from the sample, using a single transferrable voting system. The system operates as follows.

Let N be the entire set of filtered property transactions from Stage 1. Let n be the set of properties transacted in the current month. Each property in N votes for the nearest property to it in the set n.

If any property in n exceeds the threshold for election of $\frac{|N|}{0.9|n|}$ votes, then it is elected from the set; any excess votes are redistributed to its nearest neighbour. Subsequently, the property with the least number of votes is eliminated and its votes are redistributed in a similar manner. The process continues until all properties in n have either been eliminated or elected. In the end, 90% of the properties in n will be elected. This algorithm ensures there will be roughly the same number of properties included from each geographic location. In addition, because the same kinds of houses tend to be located beside each other (e.g. detached bungalows, three-bed semi-detached, apartments), the algorithm should also ensure a representative quantity of each type of dwelling.

The average monthly change in momentum of the adjusted-filtered median index was lower again, at 4.47%.

5.3. Stage 3: Localised stratification

Mix-adjustment alone is not sufficient for maximising stability between months. The reliance on a median ignores all information about the distribution below and above the median value, effectively ignoring the shape of the distribution. If this shape varies between months, such information is overlooked, thus passing up on an opportunity to enhance stability. This kind of situation arises when different regions have different medians, and the prices in these regions are diverging.

For example, the national mix-adjusted median house price at the start of 2015 was €180,000. However, because capital cities are more expensive, a large proportion of homes in Dublin were sold for more than this value (85%). The issue that hence arises is that any fluctuations in house prices that are unique to Dublin will have little impact on the overall national median.

In contrast, regions whose median price is closest to the national median will have a disproportionate effect on influencing the national median price. These areas are contributing too much information, while other areas are contributing too little. This reduces robustness, and increases volatility between months.

Accordingly, Goh et al. (2012) take the view that disaggregation of data along geographic lines is extremely important when constructing house price indexes. Studies from the Australian housing market, for example, reveal the existence of marked geographical differences in the behaviour of house prices across metropolitan areas (see Costello, Fraser, & Groenewold, 2011; Hatzvi & Otto, 2008).

One way to address this issue is through stratification. Richards and Prasad (2008) proposed a novel stratification method and tested it on an Australian data-set. They grouped together suburbs according to the long-term average price level of dwellings in those regions, taking the equally weighted average of the medians for each stratum. This measure of price growth was found to improve substantially upon an unstratified median, and was very highly correlated with regression-based measures (see also McDonald & Smith, 2009).

One limitation with Richards and Prasad's (2008) stratification technique is that it imposes arbitrary strata. The point at which a property shifts from being in one stratum to another is completely arbitrary. There is no guarantee that the strata Prasad and Richards selected reflect the most pertinent or delineated divisions in the market. Over time, these strata might shift, with more houses being built in one region than another, or a particular area being improved due to redevelopment projects. The possibility of changing relationships between the strata is not accommodated by Richards and Prasad's (2008) approach.

Our simple solution is not to impose any arbitrary stratifications, but to derive a different local base for every single property. The algorithm proceeds as follows: Two months of transaction records are selected, a stratification-base and the current month to be evaluated. We divide each sale price in the current month by that of the closest property in the stratification-base, giving a set of ratios. We then take the median of this set. This is the stratified-adjusted-filtered median.

For example, consider a house that is sold in Donegal for €105 K in February 2015. The closest house to it sold in January 2015 for €120 K. So we turn €105 K into $.875 - 1 = -12.5\%$. Alternatively, consider a house that is sold in Dublin for €420 K in February

2015. The closest house to it sold in January 2015 for €360 K. So we turn €420 K into $1.167 - 1 = +16.7\%$. Now take the median of all the percentage change values.

Under this system, all areas contribute equally to the index, thus reducing volatility. Choosing the stratification-base by default as the previous month, the average monthly change in momentum comes out at 3.76%.

5.4. Stage 4: Multiple base-month calibration

We are not limited to using only a single month as a stratification-base: we can run the same algorithm using different historical bases. For example, we can derive the index value for January 2015 using December 2014 as the stratification-base, November 2014, October 2014 and so forth.

As an example, Table 1 displays the stratified-adjusted-filtered median price change for January 2015 using the six preceding months as stratification-bases.

We recomputed the index by calculating monthly change using every available historical stratification base and averaging them. Using multiple base-month calibration, the average absolute monthly change in momentum of the stratified-adjusted-filtered median index was lower again, at 2.83%.

Table 2 displays descriptive statistics for the time series of price changes from January 2010 to July 2015 that results following the various stages of the algorithm, with the RPPI for comparison. Note that the mean and median are based on absolute monthly change, while 'smooth' refers to the average absolute monthly change in momentum.

5.5. Comparison with CSO index

Our frugal index achieved a 'smoothness' of 2.83%, which is a slightly lower level of volatility than the CSO's RPPI index, which had a 'smoothness' of 3.35% for the same period. For example, the maximum monthly change between any consecutive months for our frugal index was –6.7% for December 2012–January 2013, while the largest jump for the RPPI was a jump of +8.1% between November and December 2012. The correlation between the monthly changes of the two indexes was only $r = .43$, suggesting that they contribute slightly different sources of information.

Quigley (1995) found that hybrid models which combine information from repeat-sales and hedonic regression can be even more robust than either method in isolation. Our findings support this idea: when the two indexes are optimally weighted to minimise the smoothness value of the resulting composite index (56.1% for the frugal index, 43.9% for the RPPI), the resulting monthly change in momentum drops to only 2.51%. For the sake of comparison, the average monthly change in momentum of the RPPI's 3-month rolling average is .76%, while that of the 12-month rolling average is .25%.

Figure 1 plots the two time series, frugal and RPPI, plus their minimised volatility composition. Although our frugal index is more robust than the RPPI produced by the

Table 1. Price change between Dec 2014 and Jan 2015 using different stratification-base months.

	Jul 14	Aug 14	Sep 14	Oct 14	Nov 14	Dec 14
Change Jan 15	+5.7%	+3.9%	+3.8%	+2.7%	+2.2%	+6.0%

Table 2. Descriptive statistics for price change index produced at various stages.

	Mean (%)	Median (%)	Max(%)	Min (%)	StDev (%)	Smooth (%)
Raw average	7.01	5.74	+31.1	−17.9	9.23	12.40
Raw median	5.06	4.07	+23.8	−15.2	6.79	8.42
Stage 1	3.85	3.23	+11.1	−15.6	4.81	6.37
Stage 2	2.58	2.70	+9.19	−7.38	3.31	4.47
Stage 3	2.72	2.22	+7.33	−8.38	3.38	3.76
Stage 4	2.05	1.64	+5.41	−6.67	2.55	2.83
RPPI	2.16	1.61	+8.06	−5.50	2.73	3.35

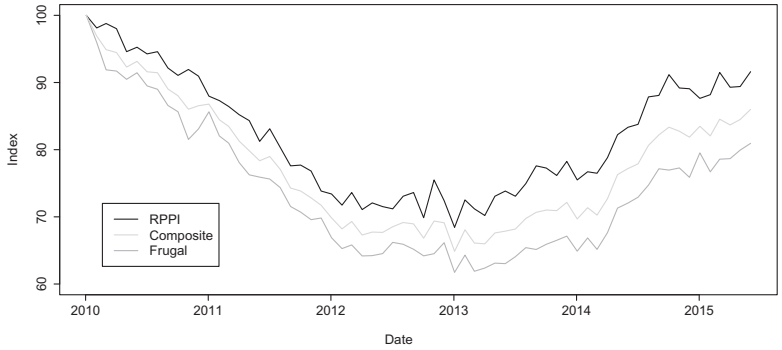


Figure 1. RPPI, frugal and composite indexes from January 2010 to June 2015.

CSO, the composite index is the most robust of all, and is what investors would choose to hold if both indexes were available to trade in an open market. By splitting their investment 56–44, investors would maximise the risk to return profile of their portfolio, and create a more robust index in the process.

6. Conclusion

We have shown that, contrary to the assertions of O'Hanlon (2011), the frugal data available from stamp duty returns, namely sale price, date of sale and address, are sufficient for developing an index that matches and exceeds the robustness of the CSO's RPPI, which relies on recording a multitude of characteristics for each property.

Admittedly, our frugal index doesn't improve greatly on the existing RPPI (though further refinements may lead to enhanced performance). The main advantage of our novel algorithm is the ease and flexibility with which it can be implemented. The code can be run on any database containing property prices and locations. It automatically controls for outliers, noise and data-set bias. As soon as new data become available, the index can be recomputed instantly with no overhead. It can also be applied to houses that have not been sold yet, using their asking prices to anticipate future changes in sale price. Because the algorithm is completely automated, it also allows users to analyse changes for any subset of records (e.g. by province, county or any arbitrarily selected geographical area).

The speedy measurement of changes in house prices is of great importance to policy-makers and investors, and is also crucial to understanding the operation of the housing market. Empirical evidence suggests that mobilising real estate derivative markets brings about significant economic benefits in the form of rapid adjustments towards supply–demand equilibriums in housing markets, lower rents on real estate and reduced amplitude of speculative house price movements (Englund et al., 2002; Lacoviello & Ortalo-Magné, 2003; Quigley, 1999). Our algorithm could be used to support derivative markets by providing an objective means of deciding a target outcome to be speculated on, one which can be recomputed hour by hour.

Critics of our frugal approach may argue that, over the period of decades, carefully calibrated statistical techniques provide a clearer picture of gradual changes in the market. This may well be the case. However, it can also be argued that an important goal of a house price index is to communicate immediate changes in market sentiment. According to Wang and Zorn (1997), there is little value in pursuing a goal of statistical or modelling accuracy if it does not lead to improved decision-making and better economic outcomes. Short- and medium-term price fluctuations can have significant impact on government and market participants, as reflected by frequent media headlines (e.g. are prices currently rising? has the market bottomed out? is this the right time to buy?) We have provided a proof of concept that algorithms using sparse and frugal data can fill this niche, providing market participants with reliable up-to-date information on house price fluctuations.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Phil Maguire is a lecturer in Computer Science in the National University of Ireland, Maynooth. His research explores portfolio optimisation, predictive modelling and the foundations of measurement.

Robert Miller is an undergraduate student at National University of Ireland, Maynooth taking the degree in Computational Thinking. He has been awarded the Alan Turing and Stephen Cook prizes in Computational Thinking and the Delort and McMahon prizes in mathematics.

Philippe Moser is a lecturer in Computer Science in the National University of Ireland, Maynooth. His research interests include algorithmic information theory, randomness, computability, complexity theory and computational finance.

Rebecca Maguire is a lecturer in Psychology in the National College of Ireland. Her area of expertise is the cognitive modelling of uncertainty.

References

Agnello, A., & Schuknecht, L. (2011). Booms and busts in housing markets: Determinants and implications. *Journal of Housing Economics, 20*, 171–190.

Bailey, M., Muth, R., & Nourse, H. (1963). A regression method for real estate price index construction. *Journal of the American Statistical Association, 58*, 933–942.

Baroni, M., Barthélémy, F., & Mokrane, M. (2005). Real estate prices: A Paris repeat sales residential index. *Journal of Real Estate Literature, 13*, 303–322.

Bourassa, S., Hoesli, M., & Sun, J. (2006). A simple alternative house price index method. *Journal of Housing Economics, 15*, 80–97.
Case, B., Pollakowski, H., & Wachter, S. (1991). On choosing among house price index methodologies. *Real Estate Economics, 19*, 286–307.
Case, B., Pollakowski, H., & Wachter, S. (1997). Frequency of transaction and house price modeling. *The Journal of Real Estate Finance and Economics, 14*, 173–187.
Case, K., Quigley, J., & Shiller, R. (2005). Comparing wealth effects: The stock market vs. the housing market. *Advances in Macroeconomics, 5*. Retrieved from https://www.degruyter.com/view/j/bejm.2005.5.1/bejm.2005.5.1.1235/bejm.2005.5.1.1235.xml
Case, K. E., & Shiller, R. J. (1987, September/October). Prices of single-family homes since 1970: New indexes for four cities. *New England Economic Review,* 45–56.
Case, B., & Szymanoski, E. J. (1995). Precision in house price indices: Findings of a comparative study of house price index methods. *Journal of Housing Research, 6*, 483–496.
Chandler, D., & Disney, R. (2014). *Measuring house prices: A comparison of difference indices*. Institute for Fiscal Studies, Briefing Note BN146. Retrieved from http://www.ifs.org.uk/bns/bn146.pdf
Choueifaty, Y., Froidure, T., & Reynier, J. (2013). Properties of the most diversified portfolio. *Journal of Investment Strategies, 2*, 49–70.
Costello, G. (2000). Pricing size effects in housing markets. *Journal of Property Research, 17*, 203–219.
Costello, G., Fraser, P., & Groenewold, N. (2011). House prices, non-fundamental components and interstate spillovers: The Australian experience. *Journal of Banking & Finance, 35*, 653–669.
Costello, G., & Watkins, C. (2002). Towards a system of local house price indices. *Housing Studies, 17*, 857–873.
Dalton, P., & Moore, K. (2014). *How to quickly adapt to new policy needs? The experience of the central statistics office, Ireland in developing house price indicators*. Cork: Central Statistics Office.
de Haan, J., & Diewert, W. E. (2011). *Handbook on residential property price indexes*. Luxembourg: Eurostat.
de Vries, P., de Haan, J., van der Wal, E., & Marin, G. (2009). A house price index based on the SPAR method. *Journal of Housing Economics, 18*, 214–223.
Diewert, W. E., & Hendriks, R. (2011). The decomposition of a house price index into land and structures components: A hedonic regression approach. *The Valuation Journal, 6*, 58–105.
Dombrow, J., Knight, J., & Sirmans, C. (1997). Aggregation bias in repeat-sales indices. *The Journal of Real Estate Finance and Economics, 14*, 75–88.
Englund, P., Hwang, M., & Quigley, J. (2002). Hedging housing risk. *The Journal of Real Estate Finance and Economics, 24*, 167–200.
Forni, M., Hallin, M., Lippi, M., & Reichlin, L. (2003). Do financial variables help forecasting inflation and real activity in the euro area? *Journal of Monetary Economics, 50*, 1243–1255.
Gillen, K., Thibodeau, T. G., & Wachter, S. (2001). Anisotropic autocorrelation in house prices. *Journal of Real Estate Finance and Economics, 23*, 5–30.
Goh, Y., Costello, G., & Schwann, G. (2012). Accuracy and robustness of house price index methods. *Housing Studies, 27*, 643–666.
Gupta, R., & Hartley, F. (2013). The role of asset prices in forecasting inflation and output in South Africa. *Journal of Emerging Market Finance, 12*, 239–291.
Gupta, R., & Kabundi, A. (2010). Forecasting macroeconomic variables in a small open economy: A comparison between small- and large-scale models. *Journal of Forecasting, 29*, 168–185.
Hamid, A. M. (2001). *Incorporating a geographic information system in hedonic modelling of farm property values* (PhD ed.). Lincoln University, Christchurch.
Hansen, J. (2009). Australian house prices: A comparison of hedonic and repeat-sales measures. *Economic Record, 85*, 132–145.
Hatzvi, E., & Otto, G. (2008). Prices, rents and rational speculative bubbles in the Sydney housing market. *Economic Record, 84*, 405–420.
Heene, M., Coyne, J., Francis, G., Maguire, P., & Maguire, R. (2014). *Crisis in cognitive science? Rise of the undead theories*. Proceedings of the 36th Annual Meeting of the Cognitive Science Society.

Ismail, S. (2006). Spatial autocorrelation and real estate studies: A literature review. *Malaysian Journal of Real Estate, 1*, 1–13.
Kain, J., & Quigley, J. (1970). Measuring the value of housing quality. *Journal of the American Statistical Association, 65*, 532–548.
Lacoviello, M., & Ortalo-Magné, F. (2003). Hedging housing risk in London. *The Journal of Real Estate Finance and Economics, 27*, 191–209.
Leishman, C. (2009). Spatial change and the structure of urban housing sub-markets. *Housing Studies, 24*, 563–585.
Lyons, R. C. (2015). East, West, boom and bust: The spread of house prices and rents in Ireland 2007–2012. *Journal of Property Research, 32*, 77–101.
Maguire, P., Kelly, S., Moser, P., & Maguire, R. (in press). Further evidence in support of the low volatility anomaly: Optimizing buy-and-hold portfolios using historical volatility. *Journal of Asset Management*.
Maguire, P., Moser, P., O'Reilly, K., McMenamin, C., Kelly, R., & Maguire, R. (2014). *Maximizing positive portfolio diversification*. IEEE Computational Intelligence for financial Engineering & Economics (CIFEr) Conference, London, 174–181.
McDonald, C., & Smith, M. (2009). *Developing stratified housing price measures for New Zealand*. Wellington: Reserve Bank of New Zealand, DP2009/07.
Munro, M. (1987). Intra-urban changes in housing prices: Glasgow 1972–1983. *Housing Studies, 2*, 65–81.
O'Hanlon, N. (2011). Constructing a national house price index for Ireland. *Statistical and Social Inquiry Society of Ireland, 40*, 167–196.
Plakandaras, V., Gupta, R., Gogas, P., & Papadimitriou, T. (2015). Forecasting the U.S. real house price index. *Economic Modelling, 45*, 259–267.
Quigley, J. M. (1995). A simple hybrid model for estimating real estate price indexes. *Journal of Housing Economics, 4*, 1–12.
Quigley, J. M. (1999). Real estate price and economic cycles. *International Real Estate Review, 2*, 1–20.
Richards, A., & Prasad, N. (2008). Improving median housing price indexes through stratification. *Journal of Real Estate Research, 30*, 45–75.
Schwann, G. M. (1998). A real estate price index for thin markets. *The Journal of Real Estate Finance and Economics, 16*, 269–287.
Shiller, R. (2003). From efficient markets theory to behavioral finance. *Journal of Economic Perspectives, 17*, 83–104.
Shimizu, C., Nishimura, K. G., & Watanabe, T. (2010). Housing prices in Tokyo: A comparison of hedonic and repeat sales measures. *Jahrbücher für Nationalökonomie und Statistik* [Journal of Economics and Statistics], *230*, 792–813.
Sommervoll, D. (2006). Temporal aggregation in repeated sales models. *The Journal of Real Estate Finance and Economics., 33*, 151–165.
Stock, J., & Watson, M. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, *23*, 405–430.
Wallace, N., & Meese, R. (1997). The construction of residential housing price indices: A comparison of repeat-sales, hedonic-regression, and hybrid approaches. *The Journal of Real Estate Finance and Economics, 14*, 51–73.
Wang, F. T., & Zorn, P. M. (1997). Estimating house price growth with repeat sales data: What's the aim of the game? *Journal of Housing Economics, 6*, 93–118.

Appendix C

Publication: GeoTree: a data structure for constant time geospatial search enabling a real-time property index

GeoTree: a data structure for constant time geospatial search enabling a real-time property index

Robert Miller and Phil Maguire

National University of Ireland, Maynooth,
Kildare, Ireland
robert.miller@mu.ie
phil.maguire@mu.ie

Abstract. A common problem appearing across the field of data science is k-NN (k-nearest neighbours), particularly within the context of Geographic Information Systems. In this article, we present a novel data structure, the GeoTree, which holds a collection of geohashes (string encodings of GPS co-ordinates). This enables a constant $O(1)$ time search algorithm that returns a set of geohashes surrounding a given geohash in the GeoTree, representing the approximate k-nearest neighbours of that geohash. Furthermore, the GeoTree data structure retains an $O(n)$ memory requirement. We apply the data structure to a property price index algorithm focused on price comparison with historical neighbouring sales, demonstrating an enhanced performance. The results show that this data structure allows for the development of a real-time property price index, and can be scaled to larger datasets with ease.

Keywords: GeoTree, geospatial, k-NN, data structure, price index

1 Introduction

Large scale datasets are a hot topic in computer science. Each one tends to present its own problems and intricacies [1]. The Nearest Neighbour (NN) problem is a well known and vital facet of many data mining research topics. This involves finding the nearest data point to a given point under some metric which measures the *distance* between data points. In the context of geospatial data, the NN problem often emerges in the form of geographical proximity search [2].

Real world geographic data is usually represented by a pair of GPS co-ordinates, which pinpoint any location on Earth with unlimited precision. As a result of their structure, computing the distance between pairs of points in order to find the *nearest neighbour* can be extremely slow on large datasets.

The problem often requires expansion to finding the k nearest neighbours (k-NN), which further increases the complexity by requiring a sorting of the distance matrix in order to extract a ranking of points by proximity. It is extremely computationally expensive to compute and rank these distances on large datasets [3].

A computationally cheap method of solving this problem would vastly improve the scalability of proximity based algorithms [2]. We propose a data structure which enables such cheap computation, the GeoTree, and explore its potential when applied to a real-world geospatial task.

2 Background

2.1 Naive geospatial search

The distance between two pieces of geospatial data defined using the GPS co-ordinate system is computed using the *haversine* formula [4]. If we wish to find the closest point in a dataset to any given point in a naive fashion, we must loop over the dataset and compute the haversine distance between each point and the given, fixed point. This is an $O(n)$ computation. If the distances are to be stored for later use, this also requires $O(n)$ memory consumption. Thus, if the closest point to every point in the dataset must be found, this requires an additional nested loop over the dataset, resulting in $O(n^2)$ memory and time complexity overall (assuming the distance matrix is stored). If such a computation is applied to a large dataset, such as the 147,635 property transactions used in the house price index developed by [5], an $O(n^2)$ algorithm can run extremely slowly even on powerful modern machines.

As GPS co-ordinates are multi-dimensional objects, it is difficult to prune and cut data from the search space without performing the haversine computation. With a considerable portion of big data being geospatial in nature, geospatial algorithms and data structures are coming under increased research attention, with the amount of personal location data available growing by approximately 20% year-on-year according to the *McKinsey Global Institute* [6]. As such, exploring alternative methods of representing GPS co-ordinates is necessary to make algorithmic improvements.

2.2 GeoHash

A geohash is a string encoding for GPS co-ordinates, allowing co-ordinate pairs to be represented by a single string of characters. The publicly-released encoding method was invented by Niemeyer in 2008 [7]. The algorithm works by assigning a geohash string to a square area on the earth, usually referred to as a *bucket*. Every GPS co-ordinate which falls inside that bucket will be assigned that geohash. The number of characters in a geohash is user-specified and determines the size of the bucket. The more characters in the geohash, the smaller the bucket becomes, and the greater precision the geohash can resolve to. While geohashes thus do not represent points on the globe, as there is no limit to the number of characters in a geohash, they can represent an arbitrarily small square on the globe and thus can be reduced to an exact point for practical purposes. Figure 1 demonstrates parts of the geohash grid on a section of map.

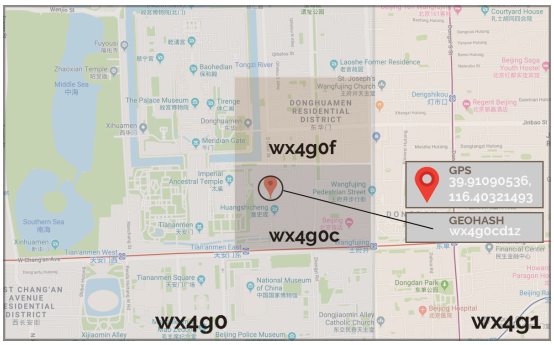


Fig. 1: GeoHash algorithm applied to a map

Geohashes are constructed in such a way that their string similarity signifies something about their proximity on the globe. Take the longest sequential substring of identical characters possible from two geohashes (starting at the first character of each geohash) and call this string x. Then x itself is a geohash (ie. a bucket) with a certain area. The longer the length of x, the smaller the area of this bucket. Thus x gives an upper bound on the distance between the points. We will refer to this substring as the *smallest common bucket* (SCB) of a pair of geohashes. We define the length of the SCB as the length of the substring defining it. This definition can additionally be generalised to a set of geohashes of any size. Furthermore, we define the SCB of a single geohash g to be the set of all geohashes in the dataset which have g as a prefix. We can immediately assert an upper bound of 123,264m for the distance between the geohashes in Figure 2, as per the table of upper bounds in the *pygeohash* package [8].

$$\text{geohash 1: } \underbrace{c_1 c_2 c_3}_{\text{SCB}} x_4 \dots x_n$$

$$\text{geohash 2: } \underbrace{c_1 c_2 c_3}_{\text{SCB}} y_4 \dots y_n$$

$$\text{where: } x_i \neq y_i \forall i \in \{4 \dots n\}$$

Fig. 2: Geohash precision example

2.3 Efficiency improvement attempts

Geohashing algorithms have, over time, improved in efficiency and have been put to use in a wide variety of applications and research contexts [9] [10]. As stated by [2], the efficient execution of nearest neighbour computations requires the use of niche spatial data structures which are constructed with the proximity of the data points being a key consideration.

The method proposed by Roussopoulos et al. [2] makes use of *R-trees*, a data structure very similar in nature to the geohash [11]. They propose an efficient algorithm for the precise *NN* computation of a spatial point, and extend this to identify the exact k-nearest neighbours using a subtree traversal algorithm which demonstrates improved efficiency over the naive search algorithm. Arya et al. [12] further this research by introducing an approximate k-NN algorithm with time complexity of $O\left(kd \log n\right)$ for any given value of k.

A comparison of some data structures for spatial searching and indexing was carried out by [13], with a specific focus on comparison between the aforementioned *R-trees* and *Quadtrees*, including application to large real-world GIS datasets. The results indicate that the Quadtree is superior to the R-tree in terms of build time due to expensive R-tree clustering. As a trade-off, the R-tree has faster query time. Both of these trees are designed to query for a very precise, user-defined area of geospatial data. As a result they are still quite slow when making a very large number of queries to the tree.

Beygelzimer et al. [14] introduce another new data structure, the cover tree. Here, each level of the tree acts as a "cover" for the level directly beneath it, which allows narrowing of the nearest neighbour search space to logarithmic time in n.

Research has also been carried out in reducing the searching overhead when the exact k-NN results are not required, and only a spatial region around each of the nearest neighbours is desired. It is often the case that ranged neighbour queries are performed as traditional k-NN queries repeated multiple times, which results in a large execution time overhead [15]. This is an inefficient method, as the lack of precision required in a ranged query can be exploited in order to optimise the search process and increase performance and efficiency, a key feature of the GeoTree.

Muja et al. provide a detailed overview of more recently proposed data structures such as partitioning trees, hashing based *NN* structures and graph based *NN* structures designed to enable efficient k-NN search algorithms [16]. The *suffix-tree*, a data structure which is designed to rapidly identify substrings in a string, has also had many incarnations and variations in the literature [17]. The GeoTree follows a somewhat similar conceptual idea and applies it to geohashes, allowing very rapid identification of groups of geohashes with shared prefixes.

The common theme within this existing body of work is the sentiment that methods of speeding up k-NN search, particularly upon data of a geospatial nature, require specialised data structures designed specifically for the purpose of proximity searching [2].

3 GeoTree

The goal of our data structure is to allow efficient approximate ranged proximity search over a set of geohashes. For example, given a database of house data, we wish to retrieve a collection of houses in a small radius around each house without having to iterate over the entire database. In more general terms, we wish to pool all other strings in a dataset which have a maximal length SCB with respect to any given string.

3.1 High-level description

A GeoTree is a general tree (a tree which has an arbitrary number of children at each node) with an immutable fixed height h set by the user upon creation. Each level of the tree represents a character in the geohash, with the exception of level zero - the root node. For example, at level one, the tree contains a node for every character that occurs among the first characters of each geohash in the database. For each node in the first level, that node will contain children corresponding to each possible character present in the second position of every geohash string in the dataset sharing the same first character as represented by the parent node. The same principle applies from level three to level h of the GeoTree, using the third to h^{th} characters of the geohash respectively.

At any node, we refer to the path to that node in the tree as the *substring* of that node, and represent it by the string where the i^{th} character corresponds to the letter associated with the node in the path at depth i.

The general structure of a GeoTree is demonstrated in Figure 3. As can be seen, the first level of the tree has a node for each possible letter in the alphabet. Only characters which are actually present in the first letters of the geohashes in our dataset will receive nodes in the constructed tree. We, however, include all characters in this diagram for clarity. In the second level, the a node also has a child for each possible letter. This same principle applies to the other nodes in the tree. Formally, at the i^{th} level, each node has a child for each of the characters present among the $(i+1)^{th}$ position of the geohash strings which are in the SCB of the current substring of that node. A worked example of a constructed GeoTree follows in Figure 4.

Consider the following set of geohashes which has been created for the purpose of demonstration: $\{gc7j98, gc7j98, gd7j98, ac7j98, gc9aaj, gc7j9d, ac7j98, gd7jya, gc9aaj\}$. The GeoTree generated by the insertion of the geohashes above with a fixed height of six would appear as seen in Figure 4.

3.2 GeoTree Data Nodes

The data attributes associated with a particular geohash are added as a child of the leaf node of the substring corresponding to that geohash in the tree, as shown in Figure 5. In the case where one geohash is associated with multiple data entries, each data entry will have its own node as a child of the geohash substring, as demonstrated in the diagram.

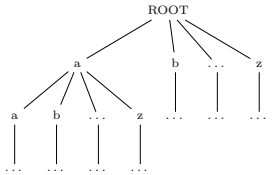


Fig. 3: GeoTree General Structure

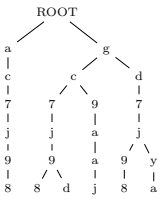


Fig. 4: Sample GeoTree Structure

It is now possible to collect all data entries in the SCB of a particular geohash substring without iterating over the entire dataset. Given a particular geohash in the tree, we can move any number of levels up the tree from that geohash's leaf nodes and explore all nearby data entries by traversing the subtree given by taking that node as the root. Thus, to compute the set of geohashes with an SCB of length m or greater with respect to the particular geohash in question, we need only explore the subtree at level m along the path corresponding to that particular geohash. Despite this improvement, we wish to remove the process of traversing the subtree altogether.

3.3 Subtree Data Caching

In order to eliminate traversal of the subtree we must cache all data entries in the subtree at each level. To cache the subtree traversal, each non-leaf node receives an additional child node which we will refer to as the *list* (*ls*) node. The list node holds references to every data entry that has a leaf node within the same subtree as the list node itself. As a result, the list node offers an instant enumeration of every leaf node within the subtree structure in which it sits, removing the need to traverse the subtree and collect the data at the leaf nodes. The structure of the tree with list nodes added is demonstrated in Figure 6 (some nodes and list nodes are omitted for the sake of brevity and clarity).

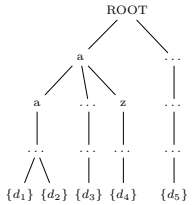


Fig. 5: GeoTree Structure with Data Nodes

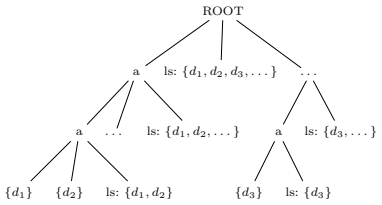


Fig. 6: GeoTree Structure with List Nodes

3.4 Retrieval of the Subtree Data

Given any geohash, we can query the tree for a set of nearby neighbouring geohashes by traversing down the GeoTree along some substring of that geohash. A longer length substring will correspond to a smaller radius in which neighbours will be returned. When the desired level is reached, the cached list node at that level can be queried for instant retrieval of the set of approximate k-NN of the geohash in question.

As a result of this structure's design, the GeoTree does not produce a distance measure for the items in the GeoTree. Rather, it clusters groups of nearby data points. While this does not allow for fine tuning of the search radius, it allows a set of data points which are geospatially close to the specified geohash to be retrieved in constant time.

3.5 Memory Requirement of the Data Structure

As each geohash is associated with only one character at each level of the GeoTree, only one node on each level will hold that geohash's data entry in its list node. Thus, each data entry is inserted into one single list node at every level of the tree. Given a tree of height h, this means that the data will be stored

in h different list nodes in addition to the one leaf node which the data receives. If the dataset is of size n, then there will be $(h+1)*n$ data entries stored in the tree. However, as the height of the tree is fixed and specified prior to the building of the tree, the overall memory requirement of the GeoTree is $O(n)$. This can be further improved to only n data entries stored by collecting a set of the data once in memory and filling the list nodes with a list of pointers to the data entries, if necessary.

3.6 Technical Implementation

To touch briefly on the implementation of GeoTree [18], a nested hash map structure is used in order to store the tree. The root node is the root hash map of the nest, with the hash keys at this level corresponding to the letters of the level one nodes. Each of these keys point to a value which is another hash map containing keys corresponding to the level two letters of geohashes which have matching first letters with the parent key. The nesting process continues down to the leaf nodes (or terminal hash values in this case) in the same fashion described in subsection 3.1. The final hash key (representing the last character of the geohash) points to the list of data entries associated with that geohash.

3.7 Time Complexity

Building (Insertion) As hash maps offer $O(1)$ insertion, insertion of data at each level of the GeoTree is $O(1)$. Furthermore, due to the height of the tree, h, being constant and fixed, insertion of entries to the GeoTree is an $O(1)$ operation overall.

SCB Lookup The $O(1)$ lookup of hash maps also means that the tree can be traversed in steps of $O(1)$ time. As the *list* nodes hold the SCB of every geohash substring possible from those in the dataset, and a maximum of h SCBs will need to be queried, it follows that any SCB lookup is also $O(1)$.

3.8 Comparison with the prefix tree (trie)

The GeoTree data structure shares a number of similarities with the prefix tree or *trie* data structure [19]. A trie is a search tree which utilises its ordering and structure to increase searching efficiency across its inserted strings. Each branch represents a character and thus as you traverse down the trie, you build the prefix of a word, working toward an entire word at each leaf node.

This is very similar to the GeoTree, as the geohash encodings of properties take the place of words in this use case and traversing the GeoTree builds prefixes of geohash strings. Both data structures make use of structure to make search more efficient, however, in the case of the GeoTree, the ordering has geographical significance rather than the semantic meaning in the prefix tree.

One key difference between tries and GeoTrees lies in the subtree data caching step. As the GeoTree relies on being able to query every entry in the subtree of a particular node, the caching is necessary to quickly return a large number of property records. In the case of a prefix tree, it would be necessary to enumerate every path in the subtree to retrieve all of the words. In the use case which the GeoTree is being applied to, this would result in a significant increase in execution time over a very large dataset.

The GeoTree data structure could be thought of as a variant or augmentation of the trie, one which is specifically designed to give a fast, approximate solution to k-NN on geospatial datasets.

4 Real-World Performance

4.1 Application: House Price Index Algorithm

In order to test the performance of GeoTree in practice, we applied it to the computation of an Irish house price index. House price indexes and forecasting models have come under increased attention from a data mining context, with a view to improve the current methods of calculating and forecasting property price changes. Such algorithms could help identify price bubbles, facilitating preemptive measures to avoid another market collapse [20,21,22].

Many of these algorithms are based around the mix-adjusted median or central price tendency model, which requires a geospatial k-NN search [5,23]. This approach is based on the principle that large amounts of aggregated data will cancel noise and result in a stable, smooth signal. It also offers the benefit of being less complex than the highly-theoretical hedonic regression model. It also requires less data than the repeat-sales model, in the sense of both quantity and time period spread [5,23,24].

Maguire et al. [5] introduced an enhanced central-price tendency model which outperformed the robustness of the hedonic regression method used by the Irish Central Statistics Office [25]. The primary limitation of this method is the algorithmic complexity and brute-force nature of the geospatial search, which impinges on its scalability to larger datasets, and restricts the introduction of further parameters. Our aim was to apply the GeoTree data structure to improve the execution time, scalability and robustness of this method. We re-implemented the algorithm used by [5] (described below), running the algorithm on the same data set (Irish Property Price Register) used in the original article as a control test for performance before introducing the GeoTree. For the purposes of algorithmic complexity calculation, we let n be the average number of house sales present in one month of the dataset, and let t be the number of months of data in the dataset.

Stage two (voting) of the **original** algorithm is executed as follows:

⇒ Iterate over each month, m, of the dataset
(t operations)
⇒ Iterate over each house, h, sold during m
(n operations)
⇒ Iterate over houses sold in m to find the nearest to h (n operations*)

Stage four (stratification) of the **original** algorithm is executed as follows:

⇒ Iterate over each month, m, of the dataset
(t operations)
⇒ Iterate over each house, h, soldHHP during m
(n operations)
⇒ Iterate over each month prior to m, m_p
($\frac{t-1}{2}$ operations)
⇒ Iterate over houses sold in m_p to find the nearest to h (n operations*)

By introducing the GeoTree to the algorithm, the steps which formerly required an $O(n)$ iteration over all houses in the dataset to identify the nearest house (marked by an asterisk) now become an $O(1)$ GeoTree ranged proximity search operation. There is, however, a mild trade-off. Rather than returning the closest property to the house in question, the GeoTree structure instead returns everything in a small area around the house (formally, it returns the maximal length non-empty SCB for that house's geohash). The bucket can then be iterated over to find the true closest property, or an alternative strategy can be employed, such as taking the median price of all houses within the small area.

4.2 Performance Results

Table 1 compares the performance of the algorithms described previously with and without GeoTrees (on a database of 279,474 property sale records), including both single threaded execution time and multi-threaded execution time (running eight threads across eight CPU cores) on our test machine. The results using the GeoTree are marked with a + symbol.

4.3 Correlation

Despite the algorithmic alteration of taking the median price of a group of geohashed nearest neighbours, as opposed to the nearest neighbour per se, the house price indexes produced by the original algorithm and the GeoTree-enhanced version are very similar. Figure 7 shows both versions of the Residential Property Price Index (RPPI) superimposed. The two different versions yielded highly correlated outputs (Pearson's $r = 0.999$, Spearman's $\rho = 0.997$, Kendall's $\tau = 0.966$), revealing that GeoTree succeeded in delivering an almost identical index to the original one, though with major performance gains in execution time.

4.4 Scalability Testing

In order to test the scalability of the GeoTree, we obtained a dataset comprising 2,857,669 property sale records for California, and evaluated both the build and query time of the data structure. Table 2 shows mean build time and mean query time on both 10% (∼285,000 records) and 100% (∼2.85 million records) of the dataset. In this context, query time refers to the total time to perform **100 sequential queries**, as a single query was too fast to accurately measure.

The results demonstrate that the height of the tree has a modest effect on the build time, while dataset size has a linear effect on build time, thus supporting the claimed $O(n)$ build time with $O(1)$ insertion. Furthermore, query time is shown to remain constant regardless of both tree height and dataset size, with negligible differences in all instances.

Table 2: Scalability Performance of GeoTree

Height h	4	5	6	7	8
Build Time (10%)[a]	**17.63s** (0.08s)	**18.10s** (0.10s)	**18.46s** (0.22s)	**18.84s** (0.08s)	**19.39s** (0.09s)
Build Time (100%)[b]	**179.67s** (0.58s)	**183.80s** (0.57s)	**183.99s** (0.52s)	**192.06s** (0.60s)	**194.31s** (0.94s)
Query Time (10%)[c]	**5.1ms** (0.3ms)	**5.2ms** (0.4ms)	**5.3ms** (0.9ms)	**5.3ms** (0.4ms)	**5.3ms** (0.5ms)
Query Time (100%)[c]	**5.4ms** (1.0ms)	**5.3ms** (0.9ms)	**5.5ms** (1.0ms)	**5.7ms** (1.3ms)	**5.6ms** (1.2ms)

[a] Build Time (10%) is the total time to insert 10% of dataset (∼285,000 records)
[b] Build Time (100%) is the total time to insert 100% of dataset (∼2.85m records)
[c] Query Time consists of total time to execute 100 sequential neighbour queries on 10% and 100% of the dataset respectively
[d] Times reported are in the format $\boldsymbol{\mu}(\sigma)$ calculated over ten trials

4.5 Discussion of Results

The results show that the GeoTree data structure offers the necessary scalability and speed of execution to expand to much larger geospatial datasets, including larger property price datasets. The biggest limitation of the GeoTree lies in

the geospatial search distance ranges being linked to the length of the geohash string encoding, thus not being alterable to any desired distance. As a result, the algorithm loses a small amount of accuracy in comparison with the original, as discussed in subsection 4.3. Despite this, the substantial gains in execution time shown in Table 1 combined with the scalability offered by an $O(1)$ search algorithm demonstrated in Table 2 make a compelling case for a worthwhile trade-off in certain applications, where execution time would become too long with exact methods, such as in the property price index application shown.

Further improvements to the algorithm which could be explored in future research include querying just the surrounding squares of a geohash grid through a GeoTree search, rather than moving up an entire level. For example, in Figure 1, a search for neighbours in *wx4g0c* which fails could explore neighbour *wx4g0f* and the other adjacent neighbouring squares before falling back to searching through the entirety of *wx4g0*. This would likely restore some of the lost accuracy previously mentioned without introducing a large execution time overhead, should a mapping of the lettering patterns be computed beforehand and used in neighbour exploration.

5 Conclusion

We have shown that the GeoTree data structure introduced in this article offers an efficient $O(1)$ method for geospatial approximate k-NN search over a collection of geohashes. The application to a real-world property price index algorithm revealed significant reductions in execution time, and potentially opens the door for a real-time property price index. The data structure also performed well when applied to a much larger dataset, demonstrating its scalability. In conclusion, any data science problem which requires geospatial sampling around a particular area can employ the GeoTree for $O(1)$ retrieval of approximate neighbours, potentially enabling, for example, fast retrieval of locations of interest to map users, or geo-targeted advertisement and social networking updates.

References

1. Hand, D.J.: Data Mining Based in part on the article "Data mining" by David Hand, which appeared in the Encyclopedia of Environmetrics. American Cancer Society (2013). DOI 10.1002/9780470057339.vad002.pub2
2. Roussopoulos, N., Kelley, S., Vincent, F.: Nearest neighbor queries. SIGMOD Rec. **24**(2), 71–79 (1995). DOI 10.1145/568271.223794. URL `http://doi.acm.org/10.1145/568271.223794`
3. Safar, M.: K nearest neighbor search in navigation systems. Mobile Information Systems **1**(3), 207–224 (2005)
4. Robusto, C.C.: The cosine-haversine formula. The American Mathematical Monthly **64**(1), 38–40 (1957). URL `http://www.jstor.org/stable/2309088`
5. Maguire, P., Miller, R., Moser, P., Maguire, R.: A robust house price index using sparse and frugal data. Journal of Property Research **33**(4), 293–308 (2016). DOI 10.1080/09599916.2016.1258718. URL `https://doi.org/10.1080/09599916.2016.1258718`

6. Lee, J.G., Kang, M.: Geospatial big data: Challenges and opportunities. Big Data Research **2**(2), 74 – 81 (2015). DOI https://doi.org/10.1016/j.bdr.2015.01.003. URL http://www.sciencedirect.com/science/article/pii/S2214579615000040. Visions on Big Data
7. Niemeyer, G.: geohash.org is public! (2008). URL https://blog.labix.org/2008/02/26/geohashorg-is-public. Accessed: 2019-05-02
8. McGinnis, W.: Pygeohash (2017). URL https://github.com/wdm0006/pygeohash. [Python]
9. Moussalli, R., Srivatsa, M., Asaad, S.: Fast and flexible conversion of geohash codes to and from latitude/longitude coordinates. In: 2015 IEEE 23rd Annual International Symposium on Field-Programmable Custom Computing Machines, pp. 179–186 (2015). DOI 10.1109/FCCM.2015.18
10. Moussalli, R., Asaad, S.W., Srivatsa, M.: Enhanced conversion between geohash codes and corresponding longitude/latitude coordinates (2015). URL https://patents.google.com/patent/US20160283515
11. Guttman, A.: R-trees: A dynamic index structure for spatial searching. SIGMOD Rec. **14**(2), 47–57 (1984). DOI 10.1145/971697.602266. URL http://doi.acm.org/10.1145/971697.602266
12. Arya, S., Mount, D.M., Netanyahu, N.S., Silverman, R., Wu, A.Y.: An optimal algorithm for approximate nearest neighbor searching fixed dimensions. J. ACM **45**(6), 891–923 (1998). DOI 10.1145/293347.293348. URL http://doi.acm.org/10.1145/293347.293348
13. Kothuri, R.K.V., Ravada, S., Abugov, D.: Quadtree and r-tree indexes in oracle spatial: a comparison using gis data. In: Proceedings of the 2002 ACM SIGMOD international conference on Management of data, pp. 546–557. ACM (2002)
14. Beygelzimer, A., Kakade, S., Langford, J.: Cover trees for nearest neighbor. In: Proceedings of the 23rd International Conference on Machine Learning, ICML '06, pp. 97–104. ACM, New York, NY, USA (2006). DOI 10.1145/1143844.1143857. URL http://doi.acm.org/10.1145/1143844.1143857
15. Bao, J., Chow, C., Mokbel, M.F., Ku, W.: Efficient evaluation of k-range nearest neighbor queries in road networks. In: 2010 Eleventh International Conference on Mobile Data Management, pp. 115–124 (2010). DOI 10.1109/MDM.2010.40
16. Muja, M., Lowe, D.G.: Scalable nearest neighbor algorithms for high dimensional data. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(11), 2227–2240 (2014). DOI 10.1109/TPAMI.2014.2321376
17. Apostolico, A., Crochemore, M., Farach-Colton, M., Galil, Z., Muthukrishnan, S.: 40 years of suffix trees. Communications of the ACM **59**(4), 66–73 (2016)
18. Miller, R.: Geotree data structure code implementation (2020). URL https://github.com/robertmiller72/GeoTree/blob/master/GeoTree.py. [Python]
19. De La Briandais, R.: File searching using variable length keys. In: Papers Presented at the the March 3-5, 1959, Western Joint Computer Conference, IRE-AIEE-ACM '59 (Western), p. 295–298. Association for Computing Machinery, New York, NY, USA (1959). DOI 10.1145/1457838.1457895. URL https://doi.org/10.1145/1457838.1457895
20. Klotz, P., Lin, T.C., Hsu, S.H.: Modeling property bubble dynamics in greece, ireland, portugal and spain. Journal of European Real Estate Research **9**(1), 52–75 (2016). DOI 10.1108/JERER-11-2014-0038. URL https://doi.org/10.1108/JERER-11-2014-0038
21. Diewert, W.E., de Haan, J., Hendriks, R.: Hedonic regressions and the decomposition of a house price index into land and structure components. Economet-

ric Reviews **34**(1-2), 106–126 (2015). DOI 10.1080/07474938.2014.944791. URL https://doi.org/10.1080/07474938.2014.944791

22. Jadevicius, A., Huston, S.: Arima modelling of lithuanian house price index. International Journal of Housing Markets and Analysis **8**(1), 135–147 (2015). DOI 10.1108/IJHMA-04-2014-0010. URL https://doi.org/10.1108/IJHMA-04-2014-0010
23. Goh, Y.M., Costello, G., Schwann, G.: Accuracy and robustness of house price index methods. Housing Studies **27**(5), 643–666 (2012). DOI 10.1080/02673037.2012.697551. URL https://doi.org/10.1080/02673037.2012.697551
24. Prasad, N., Richards, A.: Improving median housing price indexes through stratification. Journal of Real Estate Research **30**(1), 45–72 (2008). URL https://ideas.repec.org/a/jre/issued/v30n12008p45-72.html
25. O'Hanlon, N.: Constructing a national house price index for ireland. Journal of the Statistical and Social Inquiry Society of Ireland **40**, 167–196 (2011). URL http://hdl.handle.net/2262/62349

Table 1: Complexity and performance of the algorithms

Algorithm	Complexity	μ (1 core)[a]	σ[b]	μ (8 cores)[a]	σ[b]
Voting	$O\left(n^2t\right)$	233.54 seconds[c]	2.37%	46.73 seconds[c]	1.69%
Voting$^+$	$O\left(nt\right)$	12.78 seconds[c]	1.68%	3.02 seconds[c]	0.69%
Stratify	$O\left(\frac{n^2t(t-1)}{2}\right)$	29.03 hours	2.41%	4.19 hours	1.89%
Stratify$^+$	$O\left(\frac{nt(t-1)}{2}\right)$	~0.05 hours (163.89s)	1.71%	~0.01 hours (39.63s)	0.85%
Overall	$O\left(\frac{n^2t(t+1)}{2}\right)$	29.11 hours	2.43%	4.21 hours	1.90%
Overall$^+$	$O\left(\frac{nt(t+1)}{2}\right)$	~0.05 hours (177.73s)	1.67%	~0.01 hours (43.71s)	0.79%

[a] Execution times reported are the mean (μ) of ten trials.
[b] Standard deviation (σ) reported as a percentage of the mean (μ).
[c] **Includes build time** for the dataset array / GeoTree on the dataset, as applicable.
[d] All algorithms computed using an AMD Ryzen 2700X CPU.
[e] All algorithms executed on the Irish Residential Property Price Register database of **279,474 property sale records** as of time of execution.

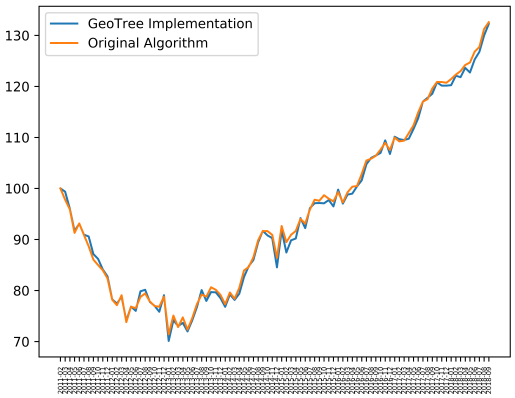


Fig. 7: Irish RPPI (GeoTree vs Original), from 02-2011 to 09-2018

Appendix D

Publication: A rapidly updating stratified mix-adjusted median property price index model

A rapidly updating stratified mix-adjusted median property price index model

Robert Miller* and Phil Maguire†
Dept. of Computer Science, National University of Ireland, Maynooth,
Kildare, Ireland.
Email: *robert.miller@mu.ie, †phil.maguire@mu.ie

***Abstract*—Homeowners, first-time buyers, banks, governments and construction companies are highly interested in following the state of the property market. Currently, property price indexes are published several months out of date and hence do not offer the up-to-date information which housing market stakeholders need in order to make informed decisions. In this article, we present an updated version of a central-price tendency based property price index which uses geospatial property data and stratification in order to compare similar houses. The expansion of the algorithm to include additional parameters owing to a new data structure implementation and a richer dataset allows for the construction of a far smoother and more robust index than the original algorithm produced.**

I. Introduction

House price indexes provide vital information to the political, financial and sales markets, affecting the operation and services of lending institutions greatly and influencing important governmental decisions [1]. As one of the largest asset classes, house prices can even offer insight regarding the overall state of the economy of a nation [2]. Property value trends can predict near-future inflation or deflation and also have a considerable effect on the gross domestic product and the financial markets [3], [4].

There are a multitude of stakeholders interested in the development and availability of an algorithm which can offer an accurate picture of the current state of the housing market, including home buyers, construction companies, governments, banks and homeowners [5], [6].

Due to the recent global financial crisis, house price indexes and forecasting models play a more crucial role than ever. The key to providing a more robust and up-to-date overview of the housing market lies in machine learning and statistical analysis on set of big data [7]. The primary aim is the improvement of currently popular algorithms for calculating and forecasting price changes, while making such indexes faster to compute and more regularly updated. Such advances could potentially play a key role in identifying price bubbles and preventing future collapses in the housing market [8], [9].

Hedging against market risk has been shown to be potentially beneficial to all stakeholders, however, it relies on having up-to-date and reliable price change information which is generally not publicly available [7], [10]. This restricts the possibility of this tool becoming a mainstream option to homeowners and small businesses.

In this article, we will expand upon previous work by [5] on a stratified, mix-adjusted median property price model by applying that algorithm to a larger and richer dataset of property listings and explore the enhancements in smoothness offered by evolving the original algorithm enabled by the use of a new data structure [11].

978-1-7281-2547-3/20/$31.00 ©2020 IEEE

II. Property price index models

In this section we will detail the three main classes of existing property price indexes. These consist of the *hedonic regression*, *repeat-sales* and *central-price tendency* methods.

A. Hedonic Regression

Hedonic regression [12] is a method which considers all of the characteristics of a house (eg. bedrooms, bathrooms, land size, location etc.) and calculates how much weight each of these attributes have in relation to the overall price of the house. While it has been shown to be the most robust measure in general by [13], outperforming the repeat-sales and mix-adjusted median methods, it requires a vast amount of detailed data and the interpretation of an experienced statistician in order to produce a result [5], [14].

As hedonic regression rests on the assumption that the price of a property can be broken down into its integral attributes, the algorithm in theory should consider every possible characteristic of the house. However, it would be impractical to obtain all of this information. As a result, specifying a complete set of regressors is extremely difficult [15].

The great number of free parameters which require tuning in hedonic regression also leads to a high chance of overfitting the model [5].

B. Repeat-sales

The repeat-sales method [16] is the most commonly used method of reporting housing sales in the United States and uses repeated sales of the same property over long periods of time to calculate change. An enhanced, weighted version of this algorithm was explored by [17]. The advantage of this method comes in the simplicity of constructing and understanding the index; historical sales of the same property are compared with each other and thus the attributes of each house need not be known nor considered. The trade-off for this simplicity comes

at the cost of requiring enormous amounts of data stretched across long periods of time [18].

It has also been theorised that the sample of repeat sales is not representative of the housing market as a whole. For example, in a study by [19], only 7% of detached homes were resold in the study period, while 30% of apartments had multiple sales in the same dataset. It is argued that this phenomenon occurs due to the 'starter home hypothesis': houses which are cheaper and in worse condition generally sell more frequently due to young homeowners upgrading [19], [20], [21]. This leads to over-representation of inexpensive and poorer quality property in the repeat-sales method. Cheap houses are also sometimes purchased for renovation or are sold quickly if the homeowner becomes unsatisfied with them, which contributes to this selection bias [19]. Furthermore, newly constructed houses are under-represented in the repeat-sales model as a brand new property cannot be a repeat sale unless it is immediately sold on to a second buyer [20].

As a result of the low number of repeat transactions, an overwhelming amount of data is discarded [22]. This leads to great inefficiency of the index and its use of the data available to it. In the commonly used repeat-sales algorithm by [17], almost 96% of the property transactions are disregarded due to incompatibility with the method [15].

C. Central Price Tendency

Central-price tendency models have been explored as an alternative to the more commonly used methods detailed previously. The model relies on the principle that large sets of clustered data tend to exhibit a noise-cancelling effect and result in a stable, smooth output [5]. Furthermore, central price tendency models offer a greater level of simplicity than the highly-theoretical hedonic regression model. When compared to the repeat sales method, central tendency models offer more efficient use of their dataset, both in the sense of quantity and time period spread [5], [23].

According to a study of house price index models by [13], the central-tendency method employed by [23] significantly outperforms the repeat-sales method despite utilising much smaller dataset. However, the method is criticised as it does not consider the constituent properties of a house and is thus more prone to inaccurate fluctuations due to a differing mix of sample properties between time periods [13]. For this reason, [13] finds that the hedonic regression model still outperforms the mix-adjusted median model used by [23]. Despite this, the simplicity and data utilisation that the method offers deserve credit were argued to justify these drawbacks [23], [13].

An enhancement to the mix-adjusted median algorithm by [23] was later shown to outperform the robustness of the hedonic regression model used by the Irish Central Statistics Office [5], [24]. The primary drawback of this algorithm was long execution time and high algorithmic complexity due to brute-force geospatial search, limiting the algorithm from being further expanded, both in terms of algorithmic features and the size of the dataset [11].

D. Improvement Attempts

With an aim to overcome the issue of algorithmic complexity in the method described by [5], a niche data structure was designed primarily for the purpose of greatly speeding up the geospatial proximity search with the aim of sacrificing minimal algorithmic precision. The *GeoTree* offers a substantial performance improvement when applied to the original algorithm while producing an almost identical index [11]. Through application of the GeoTree, the restrictions on the original algorithm have been lifted and we can now explore the performance of an evolved implementation of the algorithm on a richer, alternative dataset while introducing further parameters.

III. Case Study: MyHome Property Listing Data

MyHome [25] are a major player in property sale listings in Ireland. With data on property asking prices being collected since 2011, MyHome have a rich database of detailed data regarding houses which have been listed for sale. MyHome have provided access to their dataset for the purposes of this research.

A. Dataset Overview

The data provided by MyHome includes verified GPS co-ordinates, the number of bedrooms, the type of dwelling and further information for most of its listings. It is important to note, however, that this dataset consists of asking prices, rather than the sale prices featured in the less detailed Irish Property Price Register Data (used in the original algorithm) [5].

The dataset consists of a total of 718,351 property listing records over the period February 2011 to March 2019 (inclusive). This results in 7,330 mean listings per month (with a standard deviation of 1,689), however, this raw data requires some filtering for errors and outliers.

B. Data Filtration

As with the majority of human collected data, some pruning must be done to the MyHome dataset in order to remove outliers and erroneous data. Firstly, not all transactions in the dataset include verified GPS co-ordinates or include data on the number of bedrooms. These records will be instantly discarded for the purpose of the enhanced version of the algorithm. They account for 16.5% of the dataset. Furthermore, any property listed with greater than six bedrooms will not be considered. These properties are not representative of a standard house on the market as the number of such listings amounts to just 1% of the entire dataset.

Any data entries which do not include an asking price cannot be used for house price index calculation and must be excluded. Such records amount to 3.6% of the dataset. Additionally, asking price records which have a price of less than €10,000 or more than €1,000,000 are also excluded, as these generally consist of data entry errors (eg. wrong number of zeroes in user-entered asking price), abandoned or dilapidated properties in listings below the lower bound and mansions or commercial property in the entries exceeding the

upper bound. Properties which meet these exclusion criteria based on their price amount to only 2% of the dataset and thus are not representative of the market overall.

In summation, 77% of the dataset survives the pruning process. This leaves us with 5,646 filtered mean listings per month.

C. Comparison with PPR Dataset

The mean number of filtered monthly listings available in our dataset represents a 157% increase on the 2,200 mean monthly records used in the original algorithm's index computation [5]. Furthermore, the dataset in question is significantly more precise and accurate than the PPR dataset, owing to the ability to more effectively prune the dataset. The PPR dataset consists of address data entered by hand from written documents and does not use the Irish postcode system, meaning that addresses are often vague or ambiguous. This results in some erroneous data being factored into the model computation as there is no effective way to prune this data [5]. The MyHome dataset has been filtered to include verified addresses only, as described previously.

The PPR dataset has no information on the number of bedrooms or any key characteristics of the property. This can result in dilapidated properties, apartment blocks, inherited properties (which have an inaccurate sale value which is used for taxation purposes) and mansions mistakenly being counted as houses [5]. Our dataset consists of only single properties and the filtration process described previously greatly reduces the number of such unrepresentative samples making their way into the index calculation.

The "sparse and frugal" PPR dataset was capable of outperforming the CSO's hedonic regression model with a mix-adjusted median model [5]. With the larger, richer and more well-pruned MyHome dataset, further algorithmic enhancements to this model are possible.

IV. Performance Measures

Property prices are generally assumed to change in a smooth, calm manner over time [26] [27]. According to [5], the smoothest index is, in practice, the most robust index. As a result of this, smoothness is considered to be one of the strong indicators of reliability for an index. However, the 'smoothness' of a time series is not well defined nor immediately intuitive to measure mathematically.

The standard deviation of the time series will offer some insight into the spread of the index around the mean index value. A high standard deviation indicates that the index changes tend to be large in magnitude. While this is useful in investigating the "calmness" of the index (how dramatic its changes tend to be), it is not a reliable smoothness measure, as it is possible to have a very smooth graph with sizeable changes.

The standard deviation of the differences is a much more reliable measure of smoothness. A high standard deviation of the differences indicates that there is a high degree of variance among the differences ie. the change from point to point is unpredictable and somewhat wild. A low value for this metric would indicate that the changes in the graph behave in a more calm manner.

Finally, we present a metric which we have defined, the *mean spike magnitude* $\mu_{\Delta X}$ (MSM) of a time series X. This is intended to measure the mean value of the contrast between changes each time the trend direction of the graph flips. In other words, it is designed to measure the average size of the 'spikes' in the graph.

Given $D_X = \{d_1, \ldots, d_n\}$ is the set of differences in the time series X, we say that the pair (d_i, d_{i+1}) is a spike if d_i and d_{i+1} have different signs. Then $S_i = |d_{i+1} - d_i|$ is the spike magnitude of the spike (d_i, d_{i+1}).

The *mean spike magnitude* of X is defined as:

$$\mu_{\Delta X} = \frac{1}{|S_X|} \sum_{S \in S_X} S^2$$

where:

$S_X = \{S_1, S_2, ..., S_t\}$ is the set of all spike magnitudes of X

V. Algorithmic Evolution

A. Original Price Index Algorithm

The central price tendency algorithm introduced by [5] was designed around a key limitation; extremely frugal data. The only data available for each property was location, sale date and sale price. The core concept of the algorithm relies on using geographical proximity in order to match similar properties historically for the purpose of comparing sale prices. While this method is likely to match certain properties inaccurately, the key concept of central price tendency is that these mismatches should average out over large datasets and cancel noise.

The first major component of the algorithm is the voting stage. The aim of this is to remove properties from the dataset which are geographically isolated. The index relies on matching historical property sales which are close in location to a property in question. As a result, isolated properties will perform poorly as it will not be possible to make sufficiently near property matches for them.

In order to filter out such properties, each property in the dataset gives one vote to its closest neighbour, or a certain, set number of nearest neighbours. Once all of these votes have been casted, the total number of votes per property is enumerated and a segment of properties with the lowest votes is removed. In the implementation of the algorithm used in [5], this amounted to ten percent of the dataset.

Once the voting stage of the algorithm is complete, the next major component is the stratification stage. This is the core of the algorithm and involves stratifying average property changes on a month by month comparative basis which then serve as multiple points of reference when computing the overall monthly change. The following is a detailed explanation of the original algorithm's implementation.

First, take a particular month in the dataset which will serve as the stratification base, m_b. Then we iterate through each house sale record in m_b, represented by h_{m_b}. We must now find the nearest neighbour of h_{m_b} in each preceding month in the dataset, through a proximity search. For each prior month m_x to m_b, refer to the nearest neighbour in m_x to h_{m_b} in question as h_{m_x}. Now we are able to compute the change between the sale price of h_{m_b} and the nearest sold neighbour to h in each of the months $\{m_1, \ldots, m_n\}$ as a ratio of h_{m_b} to h_{m_x} for $x \in \{1, \ldots, n\}$. Once this is done for every property in m_b, we will have a scenario such that there is a catalogue of sale price ratios for every month prior to m and thus we can look at the median price difference between m and each historic month.

However, this is only stratification with one base, referred to as stage three in the original article [5]. We then expand the algorithm by using every month in the dataset as a stratification base. The result of this is that every month in the dataset now has price reference points to every month which preceded it and we can now use these reference points as a way to compare month to month.

Assume that m_x and m_{x+1} are consecutive months in the dataset and thus we have two sets of median ratios $\{r_x(m_1), \ldots, r_x(m_{x-1})\}$ and $\{r_{x+1}(m_1), \ldots, r_{x+1}(m_x)\}$ where $r_a(m_y)$ represents the median property sale ratio between months m_a and m_y where m_a is the chosen stratification base. In order to compute the property price index change from m_x to m_{x+1}, we look at the difference between $r_x(m_i)$ and $r_{x+1}(m_i)$ for each $i \in 1, \ldots, x-1$ and take the mean of those differences. As such, we are not directly comparing each month, rather we are contrasting the relationship of both months in question to each historical month and taking an averaging of those comparisons.

This results in a central price tendency based property index that outperformed the national Irish hedonic regression based index while using a far more frugal set of data to do so.

B. GeoTree

The largest drawback of the original index lies in the computational complexity; it is extremely slow to run. This is due to the performance impact of requiring repeated search for neighbours to each data point. This limitation was responsible for preventing the algorithm scaling to larger datasets, more refined time periods and more regular updating. A custom data structure, the GeoTree, was developed in order to trade off a small amount of accuracy in return for the ability to retrieve a cluster of neighbours to any property in constant time [11]. This data structure relies on representing the geographical location of properties as geohash strings.

The GeoTree data structure functions by placing the geohash character by character into a tree-structure where each branch at each level represents an alphanumeric character. Under each branch of the tree there is also a list node which caches all of the property records which exist as an entry in that subtree, allowing the $O(1)$ retrieval of those records. The number of sequential characters in common from the start of a pair of geohashes puts a bound on the distance between those two geohashes. Thus, by traversing down the tree and querying the list nodes, the GeoTree can return a list of approximate nearest neighbours in $O(1)$ time [11].

As can be seen in [11, Table I], the performance improvement to the index offered by the GeoTree is profound and sacrifices very little in terms of precision, with the resulting indexes proving close to identical. This development allows the scope of the index algorithm to be widened, including the introduction of larger datasets with richer data, more frequent updating and the development of new algorithmic features, some of which will be explored in this article.

C. Geohash⁺

Extended geohashes, which we will refer to as geohash⁺, are geohashes which have been modified to encode additional information regarding the property at that location. Additional parameters are encoded by adding a character in front of the geohash. The value of the character at that position corresponds to the value of the parameter which that character represents. Figure 1 demonstrates the structure of a geohash⁺ with two additional parameters, p_1 and p_2.

Fig. 1: geohash⁺ format

Any number of parameters can be prepended to the geohash. In the context of properties, this includes the number of bedrooms, the number of bathrooms, an indicator of the type of property (detached house, semi-detached house, apartment etc.), a parameter representing floor size ranges and any other attribute desired for comparison.

Alternative applications of geohash⁺ could include a situation where a rapid survey of nearby live vehicles of a certain type is required. If we prepend a parameter to the geohash locations of vehicles representing that vehicle's type, eg: 1 for cars, 2 for vans, 3 for motorcycles and so forth, we can use the GeoTree data structure to rapidly survey the SCBs around a particular vehicle, with separate SCBs generated for each type automatically.

D. GeoTree Performance with geohash⁺

Due to the design of the GeoTree data structure, a geohash⁺ will be inserted into the tree in exactly the same manner as a regular geohash [11]. If the original GeoTree had a height of h for a dataset with h-length geohashes, then the GeoTree accepting that geohash extended to a geohash⁺ with p additional parameters prepended should have a height of $h+p$. However, both of these are fixed, constant, user-specified parameters which are independent of the number of data points, and hence do not affect the constant-time performance of the GeoTree.

The major benefit of this design is that the ranged proximity search will interpret the additional parameters as regular

geohash characters when constructing the common buckets upon insertion, and also when finding the SCB in any search, without introducing additional performance and complexity drawbacks.

E. Enhanced Price Index

In order to enhance our price index model, we prepend a parameter to the geohash of each property representing the number of bedrooms present within that property. As a result, when the GeoTree is performing the SCB computation, it will now only match properties which are both nearby and share the same number of bedrooms. This allows the index model to compare the price of properties which are more similar across the time series and thus should result in a smoother, more accurate measure of the change in prices over time.

The technical implementation of this algorithmic enhancement is handled almost entirely by the GeoTree automatically, due to its design. As described previously, the GeoTree sees the additional parameter no differently to any other character in the geohash and due to its placement at the start of the geohash, the search space will be instantly narrowed to properties with matching number of bedrooms, x, by taking the x branch in the tree at the first step of traversal.

VI. Results

We ran the algorithm on the MyHome data without factoring any additional parameters as a control step. We then created a GeoTree with geohash$^+$ entries consisting of the number of bedrooms in the house prepended to the geohash for the property.

A. Comparison of Time Series

Table I shows the performance metrics previously described applied to the algorithms discussed in this paper: Original PPR, PPR with GeoTree, MyHome without bedroom factoring and MyHome with bedroom factoring. While both the standard deviation of the differences and the MSM show that some smoothness is sacrificed by the GeoTree implementation of the PPR algorithm, the index running on MyHome's data without bedroom factoring approximately matches the smoothness of the original algorithm. Furthermore, when bedroom factoring is introduced, the algorithm produces by far the smoothest index, with the standard deviation of the differences being 26.2% lower than the PPR (original) algorithm presented in [5], while the MSM sits at 58.2% lower.

If we compare the MyHome results in isolation, we can clearly observe that the addition of bedroom matching makes a very significant impact on the index performance. While the trend of each graph is observably similar, Figure 2 demonstrates that month to month changes are less erratic and appear less prone to large, spontaneous dips. Considering the smoothness metrics, the introduction of bedroom factoring generates a decrease of 26.8% in the standard deviation of the differences and a decrease of approximately 48.4% in the MSM. These results show a clear improvement by tightening the accuracy of property matching and are promising for the potential future inclusion of additional parameters such as bedroom matching should such data become available.

Figure 2 corresponds with the results of these metrics, with the *MyHome data (bedrooms factored)* index appearing the smoothest time series of the four which are compared. It is important to note that the PPR data is based upon actual sale prices, while the MyHome data is based on listed asking prices of properties which are up for sale and as such, may produce somewhat different results.

It is a well known fact that properties sell extremely well in spring and towards the end of the year, the former being the most popular period for property sales. Furthermore, the months towards late summer and shortly after tend to be the least busy periods in the year for selling property [28]. These phenomena can be observed in Figure 2 where there is a dramatic increase in the listed asking prices of properties in the spring months and towards the end of each year, while the less popular months tend to experience a slump in price movement. As such, the two PPR graphs and the MyHome data (bedrooms not factored) graph are following more or less the same trend in price action and their graphs tend to meet often, however, the majority of the price action in the MyHome data graphs tends to wait for the popular selling months. The PPR graph does not experience these phenomena as selling property can be a long, protracted process and due to a myriad of factors such as price bidding, paperwork, legal hurdles, mortgage applications and delays in reporting, final sale notifications can happen outside of the time period in which the sale price is agreed between buyer and seller.

VII. Conclusion

The introduction of bedroom factoring as an additional parameter in the pairing of nearby properties has been shown to have a profound impact on the smoothness of the mix-adjusted median property price index, which was already shown to outperform a popularly used implementation of the hedonic regression model. This improvement was made possible due to the acquisition of a richer data set and the development of the GeoTree structure, which greatly increased the performance of the algorithm. There is future potential for the introduction of further property characteristics (such as the number of bedrooms, property type etc.) in the proximity matching part of the algorithm, should such data be acquired.

Furthermore, the design of the data structure used ensures that minimal computational complexity is added when considering the technical implementation of this algorithmic adjustment. As a result of this, the index can be computed quickly enough that it would be possible to have real-time updates (eg. up to every 5 minutes) to the price index, if a sufficiently rich stream of continuous data was available to the algorithm. Large property listing websites, such as Zillow, likely have enough *live*, incoming data that such an index would be feasible to compute at this frequency, however, this volume of data is not publicly available for testing.

TABLE I: Index Comparison Statistics

Algorithm	St. Dev	St. Dev of Differences	MSM
PPR (original)	16.524	2.191	23.30
PPR (GeoTree)	16.378	2.518	29.78
MyHome (without bedrooms)	12.898	2.209	18.91
MyHome (with bedrooms)	12.985	1.617	9.75

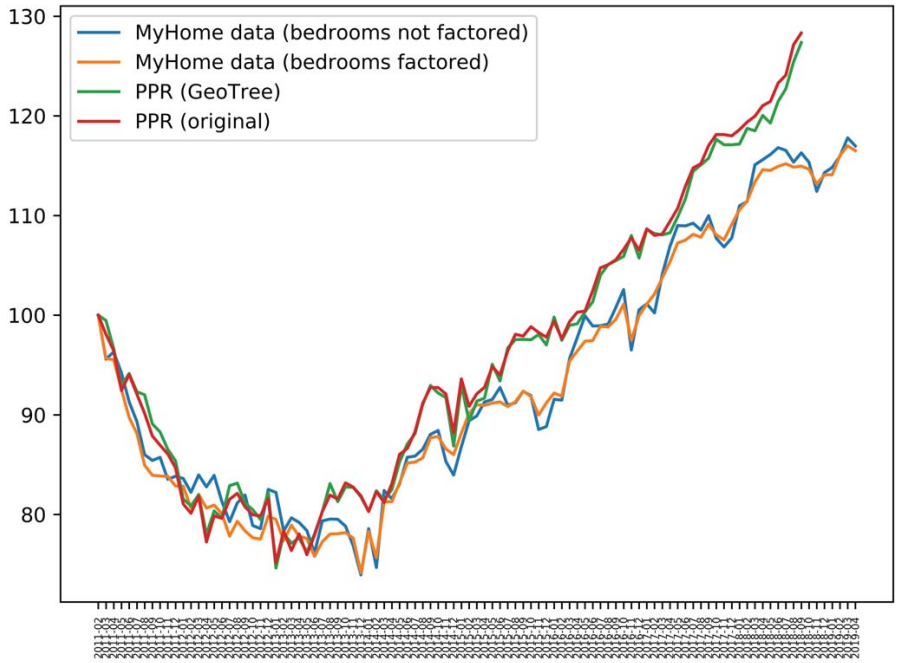


Fig. 2: Comparison of index on PPR and MyHome data sets, from 02-2011 to 03-2019 [data limited to 09-2018 for PPR]

REFERENCES

[1] W. E. Diewert, J. de Haan, and R. Hendriks, "Hedonic regressions and the decomposition of a house price index into land and structure components," *Econometric Reviews*, vol. 34, no. 1-2, pp. 106–126, 2015. [Online]. Available: https://doi.org/10.1080/07474938.2014.944791

[2] K. Case, R. Shiller, and J. Quigley, "Comparing wealth effects: The stock market versus the housing market," *Advances in Macroeconomics*, vol. 5, no. 1, 2001.

[3] M. Forni, M. Hallin, M. Lippi, and L. Reichlin, "Do financial variables help forecasting inflation and real activity in the euro area?" *Journal of Monetary Economics*, vol. 50, no. 6, pp. 1243 – 1255, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0304393203000795

[4] R. Gupta and F. Hartley, "The role of asset prices in forecasting inflation and output in south africa," *Journal of Emerging Market Finance*, vol. 12, no. 3, pp. 239–291, 2013. [Online]. Available: https://doi.org/10.1177/0972652713512913

[5] P. Maguire, R. Miller, P. Moser, and R. Maguire, "A robust house price index using sparse and frugal data," *Journal of Property Research*, vol. 33, no. 4, pp. 293–308, 2016. [Online]. Available: https://doi.org/10.1080/09599916.2016.1258718

[6] V. Plakandaras, R. Gupta, P. Gogas, and T. Papadimitriou, "Forecasting the u.s. real house price index," Rimini Centre for Economic Analysis, Working Paper series 30-14, Nov 2014. [Online]. Available: https://ideas.repec.org/p/rim/rimwps/30_14.html

[7] J. R. Hernando, *Humanizing Finance by Hedging Property Values*. Emerald Publishing Limited, 2018, ch. 10, pp. 183–204. [Online]. Available: https://www.emeraldinsight.com/doi/abs/10.1108/S0196-382120170000034015

[8] A. Jadevicius and S. Huston, "Arima modelling of lithuanian house price index," *International Journal of Housing Markets and Analysis*, vol. 8, no. 1, pp. 135–147, 2015. [Online]. Available: https://doi.org/10.1108/IJHMA-04-2014-0010

[9] P. Klotz, T. C. Lin, and S.-H. Hsu, "Modeling property bubble dynamics in greece, ireland, portugal and spain," *Journal of European Real Estate Research*, vol. 9, no. 1, pp. 52–75, 2016. [Online]. Available: https://doi.org/10.1108/JERER-11-2014-0038

[10] P. Englund, M. Hwang, and J. M. Quigley, "Hedging housing risk*," *The Journal of Real Estate Finance and Economics*, vol. 24, no. 1, pp. 167–200, Jan 2002. [Online]. Available: https://doi.org/10.1023/A:1013942607458

[11] R. Miller and P. Maguire, "GeoTree: a data structure for constant time geospatial search enabling a real-time mix-adjusted median property price index," *arXiv e-prints*, p. arXiv:2008.02167, Aug. 2020.

[12] J. F. Kain and J. M. Quigley, "Measuring the value of housing quality," *Journal of the American Statistical Association*, vol. 65, no. 330, pp. 532–548, 1970. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1970.10481102

[13] Y. M. Goh, G. Costello, and G. Schwann, "Accuracy and robustness of house price index methods," *Housing Studies*, vol. 27, no. 5, pp. 643–666, 2012. [Online]. Available: https://doi.org/10.1080/02673037.2012.697551

[14] S. Bourassa, M. Hoesli, and J. Sun, "A simple alternative house price index method," *Journal of Housing Economics*, vol. 15, no. 1, pp. 80–97, 3 2006.

[15] B. Case, H. O. Pollakowski, and S. M. Wachter, "On choosing among house price index methodologies," *Real Estate Economics*, vol. 19, no. 3, pp. 286–307, 1991. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/1540-6229.00554

[16] M. J. Bailey, R. F. Muth, and H. O. Nourse, "A regression method for real estate price index construction," *Journal of the American Statistical Association*, vol. 58, no. 304, pp. 933–942, 1963. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10480679

[17] K. E. Case and R. J. Shiller, "Prices of single family homes since 1970: New indexes for four cities," National Bureau of Economic Research, Working Paper 2393, September 1987. [Online]. Available: http://www.nber.org/papers/w2393

[18] P. de Vries, J. de Haan, E. van der Wal, and G. Mariën, "A house price index based on the spar method," *Journal of Housing Economics*, vol. 18, no. 3, pp. 214 – 223, 2009, special Issue on Owner Occupied Housing in National Accounts and Inflation Measures. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1051137709000308

[19] S. Jansen, P. Vries, H. Coolen, C. J. M. Lamain, and P. Boelhouwer, "Developing a house price index for the netherlands: A practical application of weighted repeat sales," *The Journal of Real Estate Finance and Economics*, vol. 37, pp. 163–186, 01 2008.

[20] G. COSTELLO and C. WATKINS, "Towards a system of local house price indices," *Housing Studies*, vol. 17, no. 6, pp. 857–873, 2002. [Online]. Available: https://doi.org/10.1080/02673030216001

[21] R. E. Dorsey, H. Hu, W. J. Mayer, and H. chen Wang, "Hedonic versus repeat-sales housing price indexes for measuring the recent boom-bust cycle," *Journal of Housing Economics*, vol. 19, no. 2, pp. 75 – 93, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S105113771000015X

[22] J. Dombrow, J. R. Knight, and C. F. Sirmans, "Aggregation bias in repeat-sales indices," *The Journal of Real Estate Finance and Economics*, vol. 14, no. 1, pp. 75–88, Jan 1997. [Online]. Available: https://doi.org/10.1023/A:1007720001268

[23] N. Prasad and A. Richards, "Improving median housing price indexes through stratification," *Journal of Real Estate Research*, vol. 30, no. 1, pp. 45–72, 2008. [Online]. Available: https://ideas.repec.org/a/jre/issued/v30n12008p45-72.html

[24] N. O'Hanlon, "Constructing a national house price index for ireland," *Journal of the Statistical and Social Inquiry Society of Ireland*, vol. 40, pp. 167–196, 2011. [Online]. Available: http://hdl.handle.net/2262/62349

[25] MyHomeLtd. Accessed: 2019-05-31. [Online]. Available: http://www.myhome.ie

[26] D. P. McMillen, "Neighborhood house price indexes in Chicago: a Fourier repeat sales approach," *Journal of Economic Geography*, vol. 3, no. 1, pp. 57–73, 01 2003. [Online]. Available: https://doi.org/10.1093/jeg/3.1.57

[27] J. M. Clapp, H. Kim, and A. E. Gelfand, "Predicting spatial patterns of house prices using lpr and bayesian smoothing," *Real Estate Economics*, vol. 30, no. 4, pp. 505–532, 2002. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/1540-6229.00048

[28] L. Paci, M. A. Beamonte, A. E. Gelfand, P. Gargallo, and M. Salvador, "Analysis of residential property sales using space–time point patterns," *Spatial Statistics*, vol. 21, pp. 149 – 165, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2211675317300143

Appendix E

Publication: A real-time mix-adjusted median property price index enabled by an efficient nearest neighbour approximation data structure

Journal of Banking and Financial Technology (2022) 6:135–148
https://doi.org/10.1007/s42786-022-00043-y

ORIGINAL ARTICLE

A real-time mix-adjusted median property price index enabled by an efficient nearest neighbour approximation data structure

Robert Miller[1] · **Phil Maguire**[1]

Received: 1 April 2021 / Accepted: 13 July 2022 / Published online: 24 August 2022

Abstract
Homeowners, first-time buyers, banks, governments and construction companies are highly interested in following the state of the property market. Currently, property price indexes are published several months out of date and hence do not offer the up-to-date information which housing market stakeholders need in order to make informed decisions. In this article, we present an enhanced version of a mix-adjusted median based property price index which uses geospatial property data and stratification in order to compare similar houses sold in different trading periods. The expansion of the algorithm to include additional parameters, enabled by both a richer dataset and the introduction of a new, efficient data structure for nearest neighbour approximation, allows for the construction of a far smoother and more robust index than the original algorithm produced.

Keywords House price index · Property price index · Mmix-adjusted · Property prices · House prices · Financial markets · Mortgage lenders · Inflation · Geospatial data · Geospatial index · Stratification · National statistics · Housing statistics · Property statistics · Tree structure

1 Introduction

House price indexes provide vital information to the political, financial and sales markets, affecting the operation and services of lending institutions greatly and influencing important governmental decisions [1]. As one of the largest asset classes, house prices can even offer insight regarding the overall state of the economy of a nation [2]. Property value trends can predict near-future inflation or deflation and also have a considerable effect on the gross domestic product and the financial markets [3, 4].

There are a multitude of stakeholders interested in the development and availability of an algorithm which can offer an accurate picture of the current state of the housing market, including home buyers, construction companies, governments, banks and homeowners [5, 6].

✉ Robert Miller
robert.miller@outlook.ie

Phil Maguire
phil.maguire@mu.ie

[1] Department of Computer Science, National University of Ireland, Maynooth, Kildare, Ireland

Due to the recent global financial crisis, house price indexes and forecasting models play a more crucial role than ever. The key to providing a more robust and up-to-date overview of the housing market lies in machine learning and statistical analysis on set of big data [7]. The primary aim is the improvement of currently popular algorithms for calculating and forecasting price changes, while making such indexes faster to compute and more regularly updated. Such advances could potentially play a key role in identifying price bubbles and preventing future collapses in the housing market [8, 9].

Hedging against market risk has been shown to be potentially beneficial to all stakeholders, however, it relies on having up-to-date and reliable price change information which is generally not publicly available [7, 10]. This restricts the possibility of this tool becoming a mainstream option to homeowners and small businesses.

In this article, we will expand upon previous work by Maguire et al. [5] on a stratified, mix-adjusted median property price model by applying said algorithm to a larger and richer dataset of property listings and explore the enhancements in smoothness offered by evolving the original algorithm [11]. Such evolutions have been made possible through the introduction of a custom-tailored data structure,

the GeoTree, which allows for rapid identification of a group of neighbours of any given property within fixed distance buckets [12]. These introductions remove the time barrier which was constraining the original algorithm, which took an excessive amount of time to compute results, leading to an inability to further expand the size of the dataset.

The result of our study is a flexible, highly adaptable house price index model, which can be utilised to extract an accurate house price index from datasets of varying degrees of size and richness. The model offers the ability for a layperson to construct an up-to-date property price index using publicly available data, while allowing for stakeholders with greater data resources, such as mortgage lenders or corporations, to leverage their more descriptive information sources to achieve a higher degree of model precision.

2 An overview of property price index models

In this section we will detail the three main classes of existing property price indexes. These consist of the *hedonic regression*, *repeat-sales* and *central-price tendency/mix-adjusted median* methods.

2.1 Hedonic regression

Hedonic regression is a method which considers all of the characteristics of a house (eg. bedrooms, bathrooms, land size, location etc.) and calculates how much weight each of these attributes have in relation to the overall price of the house [13]. Mathematically, a semi-log hedonic regression model is typically used for house price index estimation [14]:

$$\log(p_x) = c + \sum_{i \in I} \beta_i \mathrm{D}_i^x + \epsilon_i^x$$

where p_x is the price of property x sold in the period of interest. c is a constant. I is the set of property attributes on which the model is being fit. β_i is the regression co-efficient associated with attribute i. D_i^x is a dummy variable, indicating the presence of characteristic i in property x, in the case of categorical attributes. For continuous attributes, this will take on the continuous value in question. ϵ_i^x is an error term for attribute i.

While hedonic regression has been shown to be the most robust measure in general by Goh et al. [15], outperforming the repeat-sales and mix-adjusted median methods, it requires a vast amount of detailed data and the interpretation of an experienced statistician in order to produce a result [5, 16].

As hedonic regression rests on the assumption that the price of a property can be broken down into its integral attributes, the algorithm in theory should consider every possible characteristic of the house. However, it would be impractical to obtain all of this information. As a result, specifying a complete set of regressors is extremely difficult [17].

The great number of free parameters which require tuning in a hedonic regression model also leads to a high chance of overfitting [5]. This issue may be more pronounced in cases where the training data is sourced from a biased sample which is not representative of the property market as a whole, rather than from a complete set of property sale transactions in a given region.

2.2 Repeat-sales

The repeat-sales method is the most commonly used method of reporting housing sales in the United States and uses repeated sales of the same property over long periods of time to calculate change [18]. An enhanced, weighted version of this algorithm was explored by Case et al. [19]. The advantage of this method comes in the simplicity of constructing and understanding the index; historical sales of the same property are compared with each other and thus the attributes of each house need not be known nor considered. The trade-off for this simplicity comes at the cost of requiring enormous amounts of data stretched across long periods of time [20]. Mathematically, the standard repeat sales model takes the form [14]:

$$\log\left(\frac{p_n^t}{p_n^s}\right) = \sum_{i \in T} \beta_i \mathrm{D}_i^n + \epsilon_i^n$$

where p_n^t is the price of property n when sold in time period t. T is the set of all time periods over which the index is measuring. β_i is the regression coefficient associated with time period i. D_i^n is a dummy variable taking value 1 where $i = t$, taking value -1 where $i = s$ and taking value 0 otherwise. ϵ_i^n is an error term.

It has also been theorised that the sample of repeat sales is not representative of the housing market as a whole. For example, in a study by Jansen et al. [21], only 7% of detached homes were resold in the study period, while 30% of apartments had multiple sales in the same dataset. It is argued that this phenomenon occurs due to the 'starter home hypothesis': houses which are cheaper and in worse condition generally sell more frequently due to young homeowners upgrading [21–23]. This leads to over-representation of inexpensive and poorer quality property in the repeat-sales method. Cheap houses are also sometimes purchased for renovation or are sold quickly if the homeowner becomes unsatisfied with them, which contributes to this selection

bias [21]. Furthermore, newly constructed houses are underrepresented in the repeat-sales model as a brand new property cannot be a repeat sale unless it is immediately sold on to a second buyer [22].

As a result of the low number of repeat transactions, an overwhelming amount of data is discarded [24]. This leads to great inefficiency of the index and its use of the data available to it. In the commonly used repeat-sales algorithm by Case et al. [19], almost 96% of the property transactions are disregarded due to incompatibility with the method [17].

2.3 Central price tendency/mix-adjusted median

Central-price tendency models have been explored as an alternative to the more commonly used methods detailed previously. The model relies on the principle that large sets of clustered data tend to exhibit a noise-cancelling effect and result in a stable, smooth output [5]. Furthermore, central price tendency models offer a greater level of simplicity than the highly-theoretical hedonic regression model. When compared to the repeat sales method, central tendency models offer more efficient use of their dataset, both in the sense of quantity and time period spread [5, 25].

According to a study of house price index models by Goh et al. [15], the central-tendency method employed by Prasad et al. [25] significantly outperforms the repeat-sales method despite utilising much smaller dataset. However, the method is criticised as it does not consider the constituent properties of a house and is thus more prone to inaccurate fluctuations due to a differing mix of sample properties between time periods [15]. For this reason, Goh et al. [15] finds that the hedonic regression model still outperforms the mix-adjusted median model used by Prasad et al. [25]. Despite this, the simplicity and high level of data utilisation that the method offers were argued to justify these drawbacks [15, 25].

An evolution of the mix-adjusted median algorithm used by Prasad et al. was later shown to outperform the robustness of the hedonic regression model used by the Irish Central Statistics Office [5, 26]. This model is described in detail in Sect. 7.1. The primary drawback of this algorithm was long execution time and high algorithmic complexity due to brute-force geospatial search, limiting the algorithm from being further expanded, both in terms of algorithmic features and the size of the dataset [12].

3 The role of price indexes in the financial sector

Property price index algorithms are of high interest to financial institutions, particularly banks who partake in mortage lending. We will outline the importance of these models to said institutions, as well as exploring the feasibility of implementing each of the models discussed in Sect. 2.

3.1 Importance of property price models to the financial services sector

While there are a multitude of stakeholders in the property market, perhaps the greatest of these is the financial services sector, due to lending in the form of mortgages. For the majority of people, a house is the most valuable asset they will own in their lifetime. Furthermore, almost one-third of British households are actively paying a mortgage on their house, which collectively forms the greatest source of debt for said group of people [27].

A change in the trend of house prices can have an extraordinary impact on the general strength or weakness of an economy. When property prices are high, homeowners feel secure in increasing both spending and borrowing, which in turn stimulates economic activity and increases bank revenues. However, when house prices are falling, homeowners can reduce their spending as they begin to fear that their debt burden from their mortgage will outsize the value of their property, thus restricting economic activity [27, 28].

Mortgages are a key source of revenue for banks and financial bodies, due to their long repayment length, which results in a considerable amount of interest accrued. However, they also pose a substantial risk for said financial institutions, as they involve the lending of a large principal which is often repaid over decades, during which the financial circumstances and stability of the borrower are not guaranteed to remain constant and indeed, are often influenced by the flux in property prices as an indicator of general economic stability. This makes it difficult to predict the number of borrowers who will struggle to meet their repayments during periods of economic downturn [29].

While an economic recession usually results in massive downward pressure on commercial property prices and the equities market, such a sharp drop tends not to be reflected as drastically in the residential property market. Rather, the number of transactions usually drops, as property owners no longer wish to sell their house for a lower sum of money than they would have received before. It is likely that such a drop in residential property sales volume is reflected in a reduction in new mortgage applications, hence resulting in a loss of revenue and profit for lenders. Furthermore, such an economic event signals reduced financial stability for borrowers and thus default rates on mortgages will rise, causing a greater amount of bad debt on the books [28, 29].

It is logical then that financial bodies are highly interested in tracking the movements in property prices, to inform their lending policies and risk assessment methods. A more bullish property market may lead to banks taking on slightly more risk, with a view that the property will appreciate and

so too will the confidence of the borrower. Conversely, a bearish market will likely result in a tightening of the lending criteria, with institutions only taking on highly financially secure borrowers who they judge to be capable of weathering the storm of further depreciation of their newly-purchased property, in a worst-case scenario [30]. They might also be interested in comparing a mortgage application to the average property price for that region, to judge whether the price is excessively expensive when balanced with the financial circumstances of the applicant.

The untimely manner in which government statistical offices tend to release information on market movements, with a lag of 1–2 months being typical, may result in key policy decisions around lending being made later than is ideal. As a result, larger financial institutions are often interested in creating their own custom house price model which delivers up-to-date information, in order to better inform their lending criteria. We will present a suitable, performant model meeting these criteria later in this article.

3.2 Viability of models for application in the banking sector

Where a bank wishes to develop their own property price index model in order to get more up-to-date market information, there are some key considerations when choosing the appropriate methodology to employ. While the repeat sales method might at first seem tempting due to the simplicity of implementation, further thought reveals that this method is unlikely to be suitable. This algorithm relies on comparing multiple sales of the exact same house over long periods of time. If a financial body is using their historical mortgage data to fit the model, it is unlikely that the past sales of any given property were conducted using mortgages taken out at the same bank by different buyers, resulting in a low match rate for what is already a wasteful method in terms of data utilisation. Furthermore, historical data stretching back over decades is generally necessary to generate a reliable result with this method, which will likely be difficult for an institution to both source and convert into a clean, rich digital format [20].

The hedonic regression model may be a viable option, as these institutions will have property characteristic data for the properties on their loan books, which is key to the performance of this algorithm. However, the main drawback of using this method is the complexity of the model. The process of creating a hedonic regression model is very theoretically intense and generally requires the work of a number of statisticians in order to implement and interpret the index on an ongoing, regular basis. Furthermore, due to the human labour associated with maintaining a hedonic regression model, as well as the reliance on rich, detailed and well filtered data, it is difficult to produce the model on a more frequent time schedule than monthly or bi-monthly, particularly when this work must be repeated on a region-by-region basis, where an institution wants more granular measures than a national model.

Overfitting is another possible avenue of concern with regard to hedonic regression indices, as mentioned in Sect. 2.1. As hedonic regression relies on having a complete view of the property market, it may adapt poorly to financial institutions who likely only have access to a biased sample of property sales which have used their own lending products as the method of payment. If a particular bank was to target the middle-class working family as their intended customer base, for example, this may lead to a bias in the type of homes which are predominantly included in the model's data pipeline, thus not accurately capturing the trend in the broader housing market, rather, only the movements in a subset of it.

Mix-adjusted median based property price index models may therefore prove the most effective option for a financial institution to implement. The main advantages of such an approach lie in the ease of implementation and flexibility to incorporate various data sources of differing densities. Firstly, a mix-adjusted median algorithm can usually be computed in an entirely automated way, without a great amount of tuning or manual processing, reducing the need for multiple statisticians to spend time constantly tweaking the model to produce a monthly release, particularly where results are being produced for a number of different cities or regions. This allows for the model to be recomputed very frequently; as often as daily or two-to-three times per week, if sufficient *live* incoming data is available for the model.

This model also does not rely on specifying a complete set of price-affecting characteristics and can work with as little as three attributes: the sale date, the address and the price. Due to this, the algorithm can use the entire property sale transaction data for greater accuracy and avoidance of overfitting, which is published publicly in most countries; for example, by the Property Services Regulatory Authority in Ireland, or by HM Land Registry in the United Kingdom. Furthermore, the flexibility of the methodology allows for additional core attributes, such as the number of rooms, to be included for greater accuracy, as we will demonstrate later in this article. This means that the institution can mix their own highly detailed mortgage data together with general, unbiased but sparsely-detailed data for property sales, in order to increase the model's perspective of the market as a whole. As a result, the mix-adjusted median model is a sensible option for large banking institutions who wish to see very regular updates on the market in order to aid them in deciding on their credit lending policies.

4 GeoTree: a data structure for rapid, approximate nearest neighbour bucket searching

In order to solve the issues surrounding the brute-force geospatial search for nearest neighbours discussed in Sect. 2.3, a fast, custom bucket solution was implemented in order to generate neighbour results for a given property in $O(1)$ time. This means that the execution time does not increase with the number of samples in the structure; it remains constant regardless of the size.

4.1 Naive geospatial search

The distance between two pieces of geospatial data defined using the GPS co-ordinate system is computed using the *haversine* formula [31]. If we wish to find the closest point in a dataset to any given point in a naive fashion, we must loop over the dataset and compute the haversine distance between each point and the given, fixed point. This is an $O(n)$ computation. If the distances are to be stored for later use, this also requires $O(n)$ memory consumption. Thus, if the closest point to every point in the dataset must be found, this requires an additional nested loop over the dataset, resulting in $O(n^2)$ memory and time complexity overall (assuming the distance matrix is stored). If such a computation is applied to a large dataset, such as the 147,635 property transactions used in the house price index developed by Maguire et al. [5], an $O(n^2)$ algorithm can run extremely slowly even on powerful modern machines.

As GPS co-ordinates are multi-dimensional objects, it is difficult to prune and cut data from the search space without performing the haversine computation. Due to this, a different approach to geospatial search will prove necessary to investigate.

4.2 GeoHash

A geohash is a string encoding for GPS co-ordinates, allowing co-ordinate pairs to be represented by a single string of characters. The publicly-released encoding method was invented by Niemeyer in 2008 [32]. The algorithm works by assigning a geohash string to a square area on the earth, usually referred to as a *bucket*. Every GPS co-ordinate which falls inside that bucket will be assigned that geohash. The number of characters in a geohash is user-specified and determines the size of the bucket. The more characters in the geohash, the smaller the bucket becomes, and the greater precision the geohash can resolve to. While geohashes thus do not represent points on the globe, as there is no limit to the number of characters in a geohash, they can represent an arbitrarily small square on the globe and thus can be reduced to an exact point for practical purposes. Figure 1 demonstrates parts of the geohash grid on a section of map.

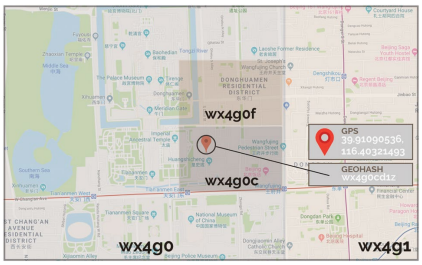


Fig. 1 GeoHash applied to a map

Fig. 2 Geohash precision example

$$\text{geohash 1: } \underbrace{c_1 c_2 c_3}_{\text{SCB}} x_4 \dots x_n$$

$$\text{geohash 2: } \underbrace{c_1 c_2 c_3}_{\text{SCB}} y_4 \dots y_n$$

$$\text{where: } x_i \neq y_i \forall i \in \{4 \dots n\}$$

Geohashes are constructed in such a way that their string similarity signifies something about their proximity on the globe. Take the longest sequential substring of identical characters possible from two geohashes (starting at the first character of each geohash) and call this string x. Then x itself is a geohash (ie. a bucket) with a certain area. The longer the length of x, the smaller the area of this bucket. Thus x gives an upper bound on the distance between the points. We will refer to this substring as the *smallest common bucket* (SCB) of a pair of geohashes. We define the length of the SCB as the length of the substring defining it. This definition can additionally be generalised to a set of geohashes of any size. Furthermore, we define the SCB of a single geohash g to be the set of all geohashes in the dataset which have g as a prefix. We can immediately assert an upper bound of 123,264 m for the distance between the geohashes in Fig. 2, as per the table of upper bounds in the *pygeohash* package, which was used in the implementation of this project [33].

4.3 Efficiency improvement attempts

Geohashing algorithms have, over time, improved in efficiency and have been put to use in a wide variety of applications and research contexts [34, 35]. As stated by Roussopoulos et al. [36], the efficient execution of nearest neighbour computations requires the use of niche spatial data structures which are constructed with the proximity of the data points being a key consideration.

The method proposed by Roussopoulos et al. [36] makes use of *R-trees*, a data structure similar in nature to the GeoTree structure which we will introduce in this article [37]. They propose an efficient algorithm for the precise *NN* computation of a spatial point, and extend this to identify the exact k-nearest neighbours using a subtree traversal algorithm which demonstrates improved efficiency over the naive search algorithm. Arya et al. [38] further this research by introducing an approximate k-NN algorithm with time complexity of $O(kd \log n)$ for any given value of k.

A comparison of some data structures for spatial searching and indexing was carried out by Kothuri et al. [39], with a specific focus on comparison between the aforementioned *R-trees* and *Quadtrees*, including application to large real-world GIS datasets. The results indicate that the Quadtree is superior to the R-tree in terms of build time due to expensive R-tree clustering. As a trade-off, the R-tree has faster query time. Both of these trees are designed to query for a very precise, user-defined area of geospatial data. As a result they are still relatively slow when making a very large number of queries to the tree.

Beygelzimer et al. [40] introduce another new data structure, the cover tree. Here, each level of the tree acts as a "cover" for the level directly beneath it, which allows narrowing of the nearest neighbour search space to logarithmic time in n.

Research has also been carried out in reducing the searching overhead when the exact k-NN results are not required, and only a spatial region around each of the nearest neighbours is desired. It is often the case that ranged neighbour queries are performed as traditional k-NN queries repeated multiple times, which results in a large execution time overhead [41]. This is an inefficient method, as the lack of precision required in a ranged query can be exploited in order to optimise the search process and increase performance and efficiency, a key feature of the GeoTree.

Muja et al. provide a detailed overview of more recently proposed data structures such as partitioning trees, hashing based *NN* structures and graph based *NN* structures designed to enable efficient k-NN search algorithms [42]. The *suffix-tree*, a data structure which is designed to rapidly identify substrings in a string, has also had many incarnations and variations in the literature [43]. The GeoTree follows a somewhat similar conceptual idea and applies it to geohashes, allowing very rapid identification of groups of geohashes with shared prefixes.

The common theme within this existing body of work is the sentiment that methods of speeding up k-NN search, particularly upon data of a geospatial nature, require specialised data structures designed specifically for the purpose of proximity searching [36].

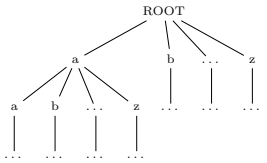


Fig. 3 GeoTree general structure

4.4 GeoTree

The goal of our data structure is to allow efficient approximate ranged proximity search over a set of geohashes. For example, given a database of house data, we wish to retrieve a collection of houses in a small radius around each house without having to iterate over the entire database. In more general terms, we wish to pool all other strings in a dataset which have a maximal length SCB with respect to any given string.

4.4.1 High-level description

A GeoTree is a general tree (a tree which has an arbitrary number of children at each node) with an immutable fixed height h set by the user upon creation. Each level of the tree represents a character in the geohash, with the exception of level zero—the root node. For example, at level one, the tree contains a node for every character that occurs among the first characters of each geohash in the database. For each node in the first level, that node will contain children corresponding to each possible character present in the second position of every geohash string in the dataset sharing the same first character as represented by the parent node. The same principle applies from level three to level h of the GeoTree, using the third to hth characters of the geohash respectively.

At any node, we refer to the path to that node in the tree as the *substring* of that node, and represent it by the string where the ith character corresponds to the letter associated with the node in the path at depth i.

The general structure of a GeoTree is demonstrated in Fig. 3. As can be seen, the first level of the tree has a node for each possible letter in the alphabet. Only characters which are actually present in the first letters of the geohashes in our dataset will receive nodes in the constructed tree, however, we include all characters in this diagram for clarity.

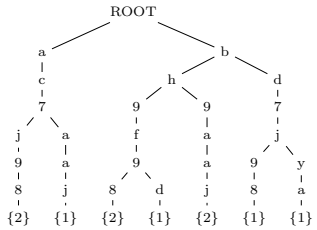


Fig. 4 Sample GeoTree structure

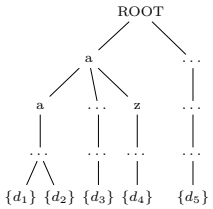


Fig. 5 GeoTree structure with data nodes

In the second level, the *a* node also has a child for each possible letter. This same principle applies to the other nodes in the tree. Formally, at the *i*th level, each node has a child for each of the characters present among the $(i + 1)$th position of the geohash strings which are in the SCB of the current substring of that node. A worked example of a constructed GeoTree follows in Fig. 4.

Consider the following set of geohashes which has been created for the purpose of demonstration: {*bh9f*98, *bh9f*98, *bd7j*98, *ac7j*98, *bh9aaj*, *bh9f*9*d*, *ac7j*98, *bd7jya*, *bh9aaj*, *ac7aaj*}. The GeoTree generated by the insertion of the geohashes above with a fixed height of six would appear as seen in Fig. 4[1].

4.4.2 GeoTree data nodes

The data attributes associated with a particular geohash are added as a child of the leaf node of the substring corresponding to that geohash in the tree, as shown in Fig. 5. In the case where one geohash is associated with multiple data entries, each data entry will have its own node as a child of the geohash substring, as demonstrated in the diagram.

[1] Note: The leaf nodes consisting of an integer in curly braces, {*x*}, is for demonstration and indicates that *x* is the number of insertions to the tree with that geohash string.

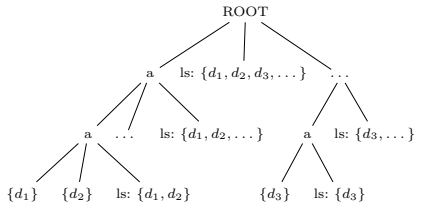
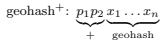


Fig. 6 GeoTree structure with list nodes

Fig. 7 geohash$^+$ format



It is now possible to collect all data entries in the SCB of a particular geohash substring without iterating over the entire dataset. Given a particular geohash in the tree, we can move any number of levels up the tree from that geohash's leaf nodes and explore all nearby data entries by traversing the subtree given by taking that node as the root. Thus, to compute the set of geohashes with an SCB of length *m* or greater with respect to the particular geohash in question, we need only explore the subtree at level *m* along the path corresponding to that particular geohash. Despite this improvement, we wish to remove the process of traversing the subtree altogether.

4.4.3 Subtree data caching

In order to eliminate traversal of the subtree we must cache all data entries in the subtree at each level. To cache the subtree traversal, each non-leaf node receives an additional child node which we will refer to as the *list* (*ls*) node. The list node holds references to every data entry that has a leaf node within the same subtree as the list node itself. As a result, the list node offers an instant enumeration of every leaf node within the subtree structure in which it sits, removing the need to traverse the subtree and collect the data at the leaf nodes. The structure of the tree with list nodes added is demonstrated in Fig. 6 (some nodes and list nodes are omitted for the sake of brevity and clarity).

4.4.4 Retrieval of the subtree data

Given any geohash, we can query the tree for a set of nearby neighbouring geohashes by traversing down the GeoTree along some substring of that geohash. A longer length substring will correspond to a smaller radius in which neighbours will be returned. When the desired level is reached, the cached list node at that level can be queried for instant

retrieval of the set of approximate k-NN of the geohash in question.

As a result of this structure's design, the GeoTree does not produce a distance measure for the items in the GeoTree. Rather, it clusters groups of nearby data points. While this does not allow for fine tuning of the search radius, it allows a set of data points which are geospatially close to the specified geohash to be retrieved in constant time, which is a worthwhile trade-off for our specific purposes, as it drastically decreases the computation time of the index.

4.5 Geohash$^+$

Extended geohashes, which we will refer to as geohash$^+$, are geohashes which have been modified to encode additional information regarding the property at that location. Additional parameters are encoded by adding a character in front of the geohash. The value of the character at that position corresponds to the value of the parameter which that character represents. Figure 7 demonstrates the structure of a geohash$^+$ with two additional parameters, p_1 and p_2.

Any number of parameters can be prepended to the geohash. In the context of properties, this includes the number of bedrooms, the number of bathrooms, an indicator of the type of property (detached house, semi-detached house, apartment etc.), a parameter representing floor size ranges and any other attribute desired for comparison.

4.6 GeoTree Performance with geohash$^+$

Due to the design of the GeoTree data structure, a geohash$^+$ will be inserted into the tree in exactly the same manner as a regular geohash [12]. If the original GeoTree had a height of h for a dataset with h-length geohashes, then the GeoTree accepting that geohash extended to a geohash$^+$ with p additional parameters prepended should have a height of $h + p$. However, both of these are fixed, constant, user-specified parameters which are independent of the number of data points, and hence do not affect the constant-time performance of the GeoTree.

The major benefit of this design is that the ranged proximity search will interpret the additional parameters as regular geohash characters when constructing the common buckets upon insertion, and also when finding the SCB in any search, without introducing additional performance and complexity drawbacks. This means that if we wish to match properties for price comparison in our index models not only geospatially, but also by bedroom count, for example, the GeoTree with geohash$^+$ will naturally take care of this added complexity without increasing the computation time complexity.

5 Case study: myhome property listing data

MyHome [44] are a major player in property sale listings in Ireland. With data on property asking prices being collected since 2011, MyHome have a rich database of detailed data regarding houses which have been listed for sale. MyHome have provided access to their dataset for the purposes of this research.

5.1 Dataset overview

The data provided by MyHome includes verified GPS co-ordinates, the number of bedrooms, the type of dwelling and further information for most of its listings. It is important to note, however, that this dataset consists of asking prices, rather than the sale prices featured in the less detailed Irish Property Price Register Data (used in the original algorithm) [5].

The dataset consists of a total of 718,351 property listing records over the period February 2011 to March 2019 (inclusive). This results in 7330 mean listings per month (with a standard deviation of 1689), however, this raw data requires some filtering for errors and outliers.

5.2 Data filtration

As with the majority of human collected data, some pruning must be done to the MyHome dataset in order to remove outliers and erroneous data. Firstly, not all transactions in the dataset include verified GPS co-ordinates or include data on the number of bedrooms. These records will be instantly discarded for the purpose of the enhanced version of the algorithm. They account for 16.5% of the dataset. Furthermore, any property listed with greater than six bedrooms will not be considered. These properties are not representative of a standard house on the market as the number of such listings amounts to just 1% of the entire dataset.

Any data entries which do not include an asking price cannot be used for house price index calculation and must be excluded. Such records amount to 3.6% of the dataset. Additionally, asking price records which have a price of less than €10,000 or more than €1,000,000 are also excluded, as these generally consist of data entry errors (eg. wrong number of zeroes in user-entered asking price), abandoned or dilapidated properties in listings below the lower bound and mansions or commercial property in the entries exceeding the upper bound. Properties which meet these exclusion criteria based on their price amount to only 2% of the dataset and thus are not representative of the market overall.

In summation, 77% of the dataset survives the pruning process. This leaves us with 5646 filtered mean listings per month.

5.3 Comparison with PPR dataset

The mean number of filtered monthly listings available in our dataset represents a 157% increase on the 2200 mean monthly records used in the original algorithm's index computation [5]. Furthermore, the dataset in question is significantly more precise and accurate than the PPR dataset, owing to the ability to more effectively prune the dataset. The PPR dataset consists of address data entered by hand from written documents and does not use the Irish postcode system, meaning that addresses are often vague or ambiguous. This results in some erroneous data being factored into the model computation as there is no effective way to prune this data [5]. The MyHome dataset has been filtered to include verified addresses only, as described previously.

The PPR dataset has no information on the number of bedrooms or any key characteristics of the property. This can result in dilapidated properties, apartment blocks, inherited properties (which have an inaccurate sale value which is used for taxation purposes) and mansions mistakenly being counted as houses [5]. Our dataset consists of only single properties and the filtration process described previously greatly reduces the number of such unrepresentative samples making their way into the index calculation.

The "sparse and frugal" PPR dataset was capable of outperforming the CSO's hedonic regression model with a mix-adjusted median model [5]. With the larger, richer and more well-pruned MyHome dataset, further algorithmic enhancements to this model are possible.

6 Performance measures

Property prices are generally assumed to change in a smooth, calm manner over time [45, 46]. According to Maguire et al. [5], the smoothest index is, in practice, the most robust index. As a result of this, smoothness is considered to be one of the strong indicators of reliability for an index. However, the 'smoothness' of a time series is not well defined nor immediately intuitive to measure mathematically.

The standard deviation of the time series will offer some insight into the spread of the index around the mean index value. A high standard deviation indicates that the index changes tend to be large in magnitude. While this is useful in investigating the "calmness" of the index (how dramatic its changes tend to be), it is not a reliable smoothness measure, as it is possible to have a very smooth graph with sizeable changes.

The standard deviation of the differences is a much more reliable measure of smoothness. A high standard deviation of the differences indicates that there is a high degree of variance among the differences ie. the change from point to point is unpredictable and somewhat wild. A low value for this metric would indicate that the changes in the graph behave in a more calm manner.

Finally, we present a metric which we have defined, the *mean spike magnitude* $\mu_{\Delta X}$ (MSM) of a time series X. This is intended to measure the mean value of the contrast between changes each time the trend direction of the graph flips. In other words, it is designed to measure the average size of the 'spikes' in the graph.

Given $D_X = \{d_1, \ldots, d_n\}$ is the set of differences in the time series X, we say that the pair (d_i, d_{i+1}) is a spike if d_i and d_{i+1} have different signs. Then $S_i = |d_{i+1} - d_i|$ is the spike magnitude of the spike (d_i, d_{i+1}).

The *mean spike magnitude* of X is defined as:

$$\mu_{\Delta X} = \frac{1}{|S_X|} \sum_{S \in S_x} S^2$$

where:

$S_X = \{S_1, S_2, \ldots, S_t\}$ is the set of all spike magnitudes of X

$|S_X|$ is the size of the set S_X

7 Algorithmic evolution

7.1 Original price index algorithm

The central price tendency algorithm introduced by Maguire et al. [5] was designed around a key limitation; extremely frugal data. The only data available for each property was location, sale date and sale price. The core concept of the algorithm relies on using geographical proximity in order to match similar properties historically for the purpose of comparing sale prices. While this method is likely to match certain properties inaccurately, the key concept of central price tendency is that these mismatches should average out over large datasets and cancel noise.

The first major component of the algorithm is the voting stage. The aim of this is to remove properties from the dataset which are geographically isolated. The index relies on matching historical property sales which are close in location to a property in question. As a result, isolated properties will perform poorly as it will not be possible to make sufficiently near property matches for them.

In order to filter out such properties, each property in the dataset gives one vote to its closest neighbour, or a certain, set number of nearest neighbours. Once all of these votes have been casted, the total number of votes per property is enumerated and a segment of properties with the lowest votes is removed. In the implementation of the algorithm used by Maguire et al. [5], this amounted to ten percent of the dataset.

Table 1 Complexity and performance of the algorithms

Algorithm	Complexity	μ (1 core)[a]	σ[b] (%)	μ (8 cores)[a]	σ[b] (%)
Voting	$O(n^2t)$	233.54 s[c]	2.37	46.73 s[c]	1.69
Voting$^+$	$O(nt)$	12.78 s[c]	1.68	3.02 s[c]	0.69
Stratify	$O\left(\frac{n^2t(t-1)}{2}\right)$	29.03 h	2.41	4.19 h	1.89
Stratify$^+$	$O\left(\frac{nt(t-1)}{2}\right)$	~0.05 h (163.89 s)	1.71	~0.01 h (39.63 s)	0.85
Overall	$O\left(\frac{n^2t(t+1)}{2}\right)$	29.11 h	2.43	4.21 h	1.90
Overall$^+$	$O\left(\frac{nt(t+1)}{2}\right)$	~0.05 h (177.73 s)	1.67	~0.01 h (43.71 s)	0.79

[a]Execution times reported are the mean (μ) of ten trials.

[b]Standard deviation (σ) reported as a percentage of the mean (μ).

[c]*Includes build time* for the dataset array/GeoTree on the dataset, as applicable.

[d]All algorithms computed using an AMD Ryzen 2700X CPU.

[e]All algorithms executed on the Irish Residential Property Price Register database of *279,474 property sale records* as of time of execution

Once the voting stage of the algorithm is complete, the next major component is the stratification stage. This is the core of the algorithm and involves stratifying average property changes on a month by month comparative basis which then serve as multiple points of reference when computing the overall monthly change. The following is a detailed explanation of the original algorithm's implementation.

First, take a particular month in the dataset which will serve as the stratification base, m_b. Then we iterate through each house sale record in m_b, represented by h_{m_b}. We must now find the nearest neighbour of h_{m_b} in each preceding month in the dataset, through a proximity search. For each prior month m_x to m_b, refer to the nearest neighbour in m_x to h_{m_b} in question as h_{m_x}. Now we are able to compute the change between the sale price of h_{m_b} and the nearest sold neighbour to h in each of the months $\{m_1, \ldots, m_n\}$ as a ratio of h_{m_b} to h_{m_x} for $x \in \{1, \ldots, n\}$. Once this is done for every property in m_b, we will have a scenario such that there is a catalogue of sale price ratios for every month prior to m and thus we can look at the median price difference between m and each historic month.

However, this is only stratification with one base, referred to as stage three in the original article [5]. We then expand the algorithm by using every month in the dataset as a stratification base. The result of this is that every month in the dataset now has price reference points to every month which preceded it and we can now use these reference points as a way to compare month to month.

Assume that m_x and m_{x+1} are consecutive months in the dataset and thus we have two sets of median ratios $\{r_x(m_1), \ldots, r_x(m_{x-1})\}$ and $\{r_{x+1}(m_1), \ldots, r_{x+1}(m_x)\}$ where $r_a(m_y)$ represents the median property sale ratio between months m_a and m_y where m_a is the chosen stratification base. In order to compute the property price index change from m_x to m_{x+1}, we look at the difference between $r_x(m_i)$ and $r_{x+1}(m_i)$ for each $i \in 1, \ldots, x-1$ and take the mean of those differences. As such, we are not directly comparing each month, rather we are contrasting the relationship of both months in question to each historical month and taking an averaging of those comparisons.

This results in a central price tendency based property index that outperformed the national Irish hedonic regression based index while using a far more frugal set of data to do so.

7.2 Enhanced price index

In order to enhance our price index model, we prepend a parameter to the geohash of each property representing the number of bedrooms present within that property. As a result, when the GeoTree is performing the SCB computation, it will now only match properties which are both nearby and share the same number of bedrooms as the property in question. This allows the index model to compare the price ratio of properties which are more similar in nature during the stratification stage and thus should result in a smoother, more accurate measure of the change in property prices over time [11].

As described previously, the GeoTree sees the additional parameter no differently to any other character in the geohash and due to its placement at the start of the geohash, the search space will be instantly narrowed to properties with matching number of bedrooms, x, by taking the x branch in the tree at the first step of traversal.

Table 2 Scalability Performance of GeoTree

Height h	4	5	6	7	8
Build time (10%)[a]	17.63 s (0.08 s)	18.10 s (0.10 s)	18.46 s (0.22 s)	18.84 s (0.08 s)	19.39 s (0.09 s)
Build time (100%)[b]	179.67 s (0.58 s)	183.80 s (0.57 s)	183.99 s (0.52 s)	192.06 s (0.60 s)	194.31 s (0.94 s)
Query time (10%)[c]	5.1 ms (0.3 ms)	5.2 ms (0.4 ms)	5.3 ms (0.9 ms)	5.3 ms (0.4 ms)	5.3 ms (0.5 ms)
Query time (100%)[c]	5.4 ms (1.0 ms)	5.3 ms (0.9 ms)	5.5 ms (1.0 ms)	5.7 ms (1.3 ms)	5.6 ms (1.2 ms)

[a]Build time (10%) is the total time to insert 10% of dataset (∼ 285,000 records)

[b]Build time (100%) is the total time to insert 100% of dataset (∼ 2.85m records)

[c]Query time consists of total time to execute 100 sequential neighbour queries on 10% and 100% of the dataset respectively

[d]Times reported are in the format $\mu(\sigma)$ calculated over ten trials

8 Results

Firstly, we will examine the performance improvement offered by the introduction of the GeoTree data structure, in addition to demonstrating the scalability of said data structure. Following this, we run the property price index algorithm on the MyHome data without factoring any additional parameters as a control step. Finally, we create a GeoTree with geohash$^+$ entries consisting of the number of bedrooms in the house prepended to the geohash for the property, showing a comparison of each index time series.

8.1 GeoTree performance

Table 1 compares the performance of the original property price index algorithm with and without use of the GeoTree (on a database of 279,474 property sale records), including both single threaded execution time and multi-threaded execution time (running eight threads across eight CPU cores) on our test machine. The results using the GeoTree are marked with a + symbol.

Figure 8 demonstrates the high level of similarity between the original *PPR* index algorithm and the *PPR* index with GeoTree. The slight difference is due to the algorithmic change of considering a basket of neighbours for each property, rather than a single neighbour per property as in the original algorithm, which could be argued to be a positive algorithmic change, as a larger sample of properties is considered. Regardless, the difference between each time series is minimal and both are extremely highly correlated with one another ($p = 0.999$).

Table 3 Index comparison statistics

Algorithm	St. dev	St. dev of differences	MSM
PPR (original)	16.524	2.191	23.30
PPR (GeoTree)	16.378	2.518	29.78
MyHome (without bedrooms)	12.898	2.209	18.91
MyHome (with bedrooms)	12.985	1.617	9.75

8.2 GeoTree scalability testing

In order to test the scalability of the GeoTree, we obtained a dataset comprising 2,857,669 property sale records for California, and evaluated both the build and query time of the data structure. Table 2 shows mean build time and mean query time on both 10% (∼ 285,000 records) and 100% (∼ 2.85 million records) of the dataset. In this context, query time refers to the total time to perform *100 sequential queries*, as a single query was too fast to accurately measure.

The results demonstrate that the height of the tree has a modest effect on the build time, while dataset size has a linear effect on build time, thus supporting the claimed $O(n)$ build time with $O(1)$ insertion. Furthermore, query time is shown to remain constant regardless of both tree height and dataset size, with negligible differences in all instances.[2]

8.3 Improved index model performance

Table 3 shows the performance metrics previously described applied to the algorithms discussed in this paper: Original PPR, PPR with GeoTree, MyHome without bedroom factoring and MyHome with bedroom factoring. While both the standard deviation of the differences and the MSM show that some smoothness is sacrificed by the GeoTree implementation of the PPR algorithm, the index running on MyHome's

[2] Note, this analysis is not designed to provide results on our house price index. Rather, it is intended to demonstrate that our GeoTree proximity matching solution is scalable to a larger geospatial dataset than the dataset used in our model analysis.

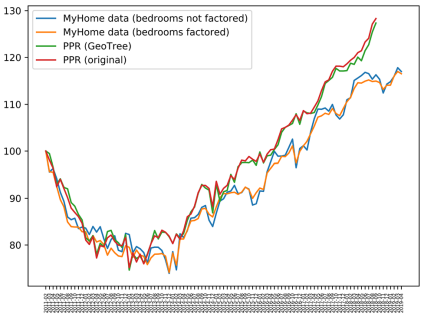


Fig. 8 Comparison of index on PPR and MyHome data sets, from 02-2011 to 03-2019 [data limited to 09-2018 for PPR]

data without bedroom factoring approximately matches the smoothness of the original algorithm. Furthermore, when bedroom factoring is introduced, the algorithm produces by far the smoothest index, with the standard deviation of the differences being 26.2% lower than the PPR (original) algorithm presented by Maguire et al. [5], while the MSM sits at 58.2% lower.

If we compare the MyHome results in isolation, we can clearly observe that the addition of bedroom matching makes a very significant impact on the index performance. While the trend of each graph is observably similar, Fig. 8 demonstrates that month to month changes are less erratic and appear less prone to large, spontaneous dips. Considering the smoothness metrics, the introduction of bedroom factoring generates a decrease of 26.8% in the standard deviation of the differences and a decrease of approximately 48.4% in the MSM. These results show a clear improvement by tightening the accuracy of property matching and are promising for the potential future inclusion of additional parameters such as bedroom matching should such data become available.

Figure 8 corresponds with the results of these metrics, with the *MyHome data (bedrooms factored)* index appearing the smoothest time series of the four which are compared. It is important to note that the PPR data is based upon actual sale prices, while the MyHome data is based on listed asking prices of properties which are up for sale and as such, may produce somewhat different results.

It is a well known fact that properties sell extremely well in spring and towards the end of the year, the former being the most popular period for property sales. Furthermore, the months towards late summer and shortly after tend to be the least busy periods in the year for selling property [47]. These phenomena can be observed in Fig. 8 where there is a dramatic increase in the listed asking prices of properties in the spring months and towards the end of each year, while the less popular months tend to experience a slump in price movement. As such, the two PPR graphs and the MyHome data (bedrooms not factored) graph are following more or less the same trend in price action and their graphs tend to meet often, however, the majority of the price action in the MyHome data graphs tends to wait for the popular selling months. The PPR graph does not experience these phenomena as selling property can be a long, protracted process and due to a myriad of factors such as price bidding, paperwork, legal hurdles, mortgage applications and delays in reporting, final sale notifications can happen outside of the time period in which the sale price is agreed between buyer and seller.

9 Conclusion

9.1 Contributions

The introduction of bedroom factoring as an additional parameter in the pairing of nearby properties has been shown to have a profound impact on the smoothness of the mix-adjusted median property price index. These developments were made possible due to the acquisition of a richer data set and the introduction of the GeoTree data structure, which greatly increased the performance of the algorithm. There is also scope for the introduction of further property characteristics (such as the number of bathrooms, property type etc.) in the proximity matching part of the algorithm, should such data be acquired.

Despite this advancement, the algorithm still has great benefit to the layperson, outperforming certain implementations of hedonic regression models without having access to richer, private datasets [11]. The result is a highly flexible algorithm, which can adapt to various levels of data availability while still offering a high degree of accuracy. Examples of free, publicly available datasets which could be used with our house price index model include the Irish Property Price Register [48] (used in this analysis), or the British Price Paid dataset covering England and Wales, published monthly by HM Land Registry [49]. Stakeholders with greater exposure to the market, such as mortgage lenders, are likely to have their own rich data sources and thus will be capable of leveraging the increased accuracy offered by our model when it is fed a more descriptive dataset, as demonstrated in this study.

Furthermore, the design of the GeoTree data structure ensures that minimal computational complexity is added when considering the technical implementation of this algorithmic adjustment [12]. Any additional parameters or attributes could also be integrated with ease, without increasing the complexity of the index computation. This contribution is of great benefit to all housing market stakeholders, as it

means that as soon as the property sale data for a given month becomes available, a house price index model can be produced immediately, using our algorithm. This counteracts the substantial time lag issue associated with national hedonic regression models, where house price indices typically are not published for one-to-two months after the end of the month to which the data pertains (e.g. August property price change may not be published until October).

9.2 Limitations and future work

The efficiency improvements offered by the GeoTree are such that our model could be computed rapidly enough, with full automation, to have real-time updates (e.g. up to every 5 min) to a property price index, if a sufficiently rich stream of continuous data was available to the algorithm. Large property listing websites, such as Zillow, likely have enough *live*, incoming listing data that such an index would be feasible to compute at this frequency, however, this volume of data is not publicly available so as to allow for demonstration of such an application by ourselves.

Comprehensive financial institutions dealing in mortgage lending likely have enough data to produce such an index on a region-by-region basis at least as frequently as weekly, if not even more regularly. This would aid their credit departments in lending decisions by offering a live, timely view of the changing dynamics of the property market and points of reference for typical house prices in each region, which the out of date national hedonic regression indices are incapable of doing, due to the lengthy publication delay discussed in Sect. 9.1.

Despite these limitations, we believe that our index has substantial benefit to all property market stakeholders, regardless of the amount of data at hand, or the richness of said data. Our future ambitions for research in this field include expanding our house price index model to perform property market forecasting based on emerging data. We also hope to gain access to an even larger property sale dataset, so that we can benchmark our model's performance on a higher frequency than monthly indices. Beyond this, our goal is to integrate the house price index model into a deep learning model framework which can perform individual property valuation based on a number of input characteristics. This model aims not only to show the present value of a given property, but also the historical change in the value of said property, using our house price index model as an input.

Funding Open Access funding provided by the IReL Consortium.

References

1. Diewert WE, de Haan J, Hendriks R (2015) Hedonic regressions and the decomposition of a house price index into land and structure components. Econom Rev 34(1–2):106–126. https://doi.org/10.1080/07474938.2014.944791
2. Case K, Shiller R, Quigley J (2001) Comparing wealth effects: the stock market versus the housing market. Adv Macroecon. https://doi.org/10.3386/w8606
3. Forni M, Hallin M, Lippi M, Reichlin L (2003) Do financial variables help forecasting inflation and real activity in the euro area? J Monetary Econ 50(6):1243–1255. https://doi.org/10.1016/S0304-3932(03)00079-5
4. Gupta R, Hartley F (2013) The role of asset prices in forecasting inflation and output in South Africa. J Emerg Market Financ 12(3):239–291. https://doi.org/10.1177/0972652713512913
5. Maguire P, Miller R, Moser P, Maguire R (2016) A robust house price index using sparse and frugal data. J Prop Res 33(4):293–308. https://doi.org/10.1080/09599916.2016.1258718
6. Plakandaras V, Gupta R, Gogas P, Papadimitriou T (2014) Forecasting the U.S. real house price index. Working paper series 30–14, Rimini Centre for Economic Analysis. https://ideas.repec.org/p/rim/rimwps/30_14.html
7. Hernando JR (2018) Humanizing finance by hedging property values. Emerald Publishing Limited, Bingley, pp 183–204. https://doi.org/10.1108/S0196-382120170000034015 (**Chap. 10**)
8. Jadevicius A, Huston S (2015) Arima modelling of Lithuanian house price index. Int J Hous Mark Anal 8(1):135–147. https://doi.org/10.1108/IJHMA-04-2014-0010
9. Klotz P, Lin TC, Hsu SH (2016) Modeling property bubble dynamics in Greece, Ireland, Portugal and Spain. J Eur Real Estate Res 9(1):52–75. https://doi.org/10.1108/JERER-11-2014-0038
10. Englund P, Hwang M, Quigley JM (2002) Hedging housing risk. J Real Estate Financ Econ 24(1):167–200. https://doi.org/10.1023/A:1013942607458
11. Miller R, Maguire P (2020) A rapidly updating stratified mix-adjusted median property price index model. In: 2020 IEEE symposium series on computational intelligence (SSCI), IEEE, p 9–15, https://doi.org/10.1109/SSCI47803.2020.9308235
12. Miller R, Maguire P (2020) GeoTree: a data structure for constant time geospatial search enabling a real-time mix-adjusted median property price index. arXiv e-prints arXiv:2008.02167
13. Kain JF, Quigley JM (1970) Measuring the value of housing quality. J Am Stat Assoc 65(330):532–548. https://doi.org/10.1080/01621459.1970.10481102
14. OECD, Eurostat, Organization IL, Fund IM, Bank TW, for Europe UNEC (2013) Handbook on residential property price indices. https://doi.org/10.1787/9789264197183-en
15. Goh YM, Costello G, Schwann G (2012) Accuracy and robustness of house price index methods. Hous Stud 27(5):643–666. https://doi.org/10.1080/02673037.2012.697551
16. Bourassa S, Hoesli M, Sun J (2006) A simple alternative house price index method. J Hous Econ 15(1):80–97. https://doi.org/10.1016/j.jhe.2006.03.001

17. Case B, Pollakowski HO, Wachter SM (1991) On choosing among house price index methodologies. Real Estate Econ 19(3):286–307. https://doi.org/10.1111/1540-6229.00554
18. Bailey MJ, Muth RF, Nourse HO (1963) A regression method for real estate price index construction. J Am Stat Assoc 58(304):933–942. https://doi.org/10.1080/01621459.1963.10480679
19. Case KE, Shiller RJ (1987) Prices of single family homes since 1970: New indexes for four cities. Working paper 2393, National Bureau of Economic Research. https://doi.org/10.3386/w2393. http://www.nber.org/papers/w2393
20. de Vries P, de Haan J, van der Wal E, Mariën G (2009) A house price index based on the spar method. J Hous Econ 18(3):214–223. https://doi.org/10.1016/j.jhe.2009.07.002 (**special Issue on Owner Occupied Housing in National Accounts and Inflation Measures**)
21. Jansen S, Vries P, Coolen H, Lamain CJM, Boelhouwer P (2008) Developing a house price index for the Netherlands: a practical application of weighted repeat sales. J Real Estate Financ Econ 37:163–186. https://doi.org/10.1007/s11146-007-9068-0
22. Costello G, Watkins C (2002) Towards a system of local house price indices. Hous Stud 17(6):857–873. https://doi.org/10.1080/02673030216001
23. Dorsey RE, Hu H, Mayer WJ, Chen Wang H (2010) Hedonic versus repeat-sales housing price indexes for measuring the recent boom-bust cycle. J Hous Econ 19(2):75–93. https://doi.org/10.1016/j.jhe.2010.04.001
24. Dombrow J, Knight JR, Sirmans CF (1997) Aggregation bias in repeat-sales indices. J Real Estate Financ Econ 14(1):75–88. https://doi.org/10.1023/A:1007720001268
25. Prasad N, Richards A (2008) Improving median housing price indexes through stratification. J Real Estate Res 30(1):45–72
26. O'Hanlon N (2011) Constructing a national house price index for Ireland. J Stat Soc Inq Soc Ireland 40:167–196
27. Bank of England (2018) How does the housing market affect the economy? https://www.bankofengland.co.uk/knowledgebank/how-does-the-housing-market-affect-the-economy. Accessed 24 May 2021
28. Zhu H et al (2005) The importance of property markets for monetary policy and financial stability. Real Estate Indica Financ Stab 21:9–29
29. Bank of England (2018) What is the bank of England's role in the housing market? https://www.bankofengland.co.uk/knowledgebank/whats-the-bank-of-englands-role-in-the-housing-market. Accessed 24 May 2021
30. Che X, Li B, Guo K, Wang J (2011) Property prices and bank lending: some evidence from China's regional financial centres. Procedia Comput Sci 4:1660–1667. https://doi.org/10.1016/j.procs.2011.04.179 (**proceedings of the International Conference on Computational Science, ICCS 2011**)
31. Robusto CC (1957) The cosine-haversine formula. Am Math Mon 64(1):38–40
32. Niemeyer G (2008) geohash.org is public! https://blog.labix.org/2008/02/26/geohashorg-is-public. Accessed 02 May 2019
33. McGinnis W (2017) Pygeohash. https://github.com/wdm0006/pygeohash, [Python]
34. Moussalli R, Srivatsa M, Asaad S (2015) Fast and flexible conversion of Geohash codes to and from latitude/longitude coordinates. In: 2015 IEEE 23rd annual international symposium on field-programmable custom computing machines, p 179–186. https://doi.org/10.1109/FCCM.2015.18
35. Moussalli R, Asaad SW, Srivatsa M (2015) Enhanced conversion between geohash codes and corresponding longitude/latitude coordinates. https://patents.google.com/patent/US20160283515
36. Roussopoulos N, Kelley S, Vincent F (1995) Nearest neighbor queries. SIGMOD Rec 24(2):71–79. https://doi.org/10.1145/568271.223794
37. Guttman A (1984) R-trees: a dynamic index structure for spatial searching. SIGMOD Rec 14(2):47–57. https://doi.org/10.1145/971697.602266
38. Arya S, Mount DM, Netanyahu NS, Silverman R, Wu AY (1998) An optimal algorithm for approximate nearest neighbor searching fixed dimensions. J ACM 45(6):891–923. https://doi.org/10.1145/293347.293348
39. Kothuri RKV, Ravada S, Abugov D (2002) Quadtree and R-tree indexes in oracle spatial: a comparison using GIS data. In: Proceedings of the 2002 ACM SIGMOD international conference on Management of data, ACM, p 546–557
40. Beygelzimer A, Kakade S, Langford J (2006) Cover trees for nearest neighbor. In: Proceedings of the 23rd international conference on machine learning, ACM, New York, NY, USA, ICML '06, p 97–104. https://doi.org/10.1145/1143844.1143857
41. Bao J, Chow C, Mokbel MF, Ku W (2010) Efficient evaluation of k-range nearest neighbor queries in road networks. In: 2010 Eleventh international conference on mobile data Management, p 115–124. https://doi.org/10.1109/MDM.2010.40
42. Muja M, Lowe DG (2014) Scalable nearest neighbor algorithms for high dimensional data. IEEE Trans Pattern Anal Mach Intell 36(11):2227–2240. https://doi.org/10.1109/TPAMI.2014.2321376
43. Apostolico A, Crochemore M, Farach-Colton M, Galil Z, Muthukrishnan S (2016) 40 years of suffix trees. Commun ACM 59(4):66–73
44. MyHome Ltd (2021) http://www.myhome.ie. Accessed 31 Mar 2021
45. McMillen DP (2003) Neighborhood house price indexes in Chicago: a Fourier repeat sales approach. J Econ Geogr 3(1):57–73. https://doi.org/10.1093/jeg/3.1.57
46. Clapp JM, Kim H, Gelfand AE (2002) Predicting spatial patterns of house prices using LPR and Bayesian smoothing. Real Estate Econ 30(4):505–532. https://doi.org/10.1111/1540-6229.00048
47. Paci L, Beamonte MA, Gelfand AE, Gargallo P, Salvador M (2017) Analysis of residential property sales using space-time point patterns. Spatial Stat 21:149–165. https://doi.org/10.1016/j.spasta.2017.06.007
48. PPR (2021) Property price register. https://www.propertypriceregister.ie. Accessed 25 Oct 2021
49. HM-Land-Registry (2021) Price paid dataset. https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads. Accessed 25 Oct 2021

Bibliography

Agnello, Luca and Ludger Schuknecht (2011). "Booms and busts in housing markets: Determinants and implications". In: *Journal of Housing Economics* 20.3, pp. 171–190.

Anderson, Hamish (July 2018). "Value of nature implicit in property prices – hedonic pricing method methodology". In: *ONS - Environmental Accounts*. URL: https://www.ons.gov.uk/economy/environmentalaccounts/methodologies/valueofnatureimplicitinpropertypriceshedonicpricingmethodhpmmethodologynote.

Anenberg, Elliot and Steven Laufer (Oct. 2017). "A More Timely House Price Index". In: *The Review of Economics and Statistics* 99.4, pp. 722–734. ISSN: 0034-6535. DOI: 10.1162/REST_a_00634.

Anselin, Luc and Nancy Lozano-Gracia (2008). "Errors in variables and spatial effects in hedonic house price models of ambient air quality". In: *Empirical economics* 34, pp. 5–34. DOI: 10.1007/s00181-007-0152-3.

Apostolico, Alberto, Maxime Crochemore, et al. (2016). "40 years of suffix trees". In: *Communications of the ACM* 59.4, pp. 66–73.

Arya, Sunil, David M. Mount, et al. (Nov. 1998). "An Optimal Algorithm for Approximate Nearest Neighbor Searching Fixed Dimensions". In: *J. ACM* 45.6, pp. 891–923. ISSN: 0004-5411. DOI: 10.1145/293347.293348.

Babyak, Michael A (2004). "What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models". In: *Psychosomatic medicine* 66.3, pp. 411–421.

Baker, Malcolm, Brendan Bradley, and Jeffrey Wurgler (2011). "Benchmarks as limits to arbitrage: Understanding the low-volatility anomaly". In: *Financial Analysts Journal* 67.1, pp. 40–54.

Bala, Alain Pholo, Dominique Peeters, and Isabelle Thomas (2014). "Spatial issues on a hedonic estimation of rents in Brussels". In: *Journal of Housing Economics* 25, pp. 104–123. ISSN: 1051-1377. DOI: 10.1016/j.jhe.2014.05.002.

Bank of England (2018a). *How does the housing market affect the economy?* Accessed: 2024-06-01. URL: https://www.bankofengland.co.uk/knowledgebank/how-does-the-housing-market-affect-the-economy.

— (2018b). *What is the Bank of England's role in the housing market?* Accessed: 2024-06-01. URL: https://www.bankofengland.co.uk/knowledgebank/whats-the-bank-of-englands-role-in-the-housing-market.

Bao, J., C. Chow, et al. (May 2010). "Efficient Evaluation of k-Range Nearest Neighbor Queries in Road Networks". In: *2010 Eleventh International Conference on Mobile Data Management*, pp. 115–124. DOI: 10.1109/MDM.2010.40.

Basu, Sabyasachi and Thomas G Thibodeau (1998). "Analysis of spatial autocorrelation in house prices". In: *The Journal of Real Estate Finance and Economics* 17, pp. 61–85. DOI: 10.1023/A:1007703229507.

Beygelzimer, Alina, Sham Kakade, and John Langford (2006). "Cover Trees for Nearest Neighbor". In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. ACM, pp. 97–104. DOI: 10.1145/1143844.1143857.

Bholat, David (2015). "Big data and central banks". In: *Big Data & Society*. DOI: 10.1177/2053951715579469.

Braun, Rahel and Sarah M Lein (2021). "Sources of bias in inflation rates and implications for inflation dynamics". In: *Journal of Money, Credit and Banking* 53.6, pp. 1553–1572. DOI: 10.1111/jmcb.12848.

Brezina, Corona (2011). *Understanding the gross domestic product and the gross national product*. Rosen Publishing Group, inc., pp. 4–5.

CACI (Mar. 2019). *Acorn technical guide - CACI*. URL: https://acorn.caci.co.uk/downloads/Acorn-Technical-document.pdf.

Cartern, Charles C. and William J. Haloupek (2000). "Spatial Autocorrelation in a Retail Context". In: *International Real Estate Review* 3.1, pp. 34–48.

Case, Bradford, Henry O. Pollakowski, and Susan M. Wachter (1991). "On Choosing Among House Price Index Methodologies". In: *Real Estate Economics* 19.3, pp. 286–307. DOI: 10.1111/1540-6229.00554.

Case, Bradford and Edward J Szymanoski (1995). "Precision in house price indices: Findings of a comparative study of house price index methods". In: *Journal of Housing Research* 6.3, pp. 483–496. URL: http://www.jstor.org/stable/24832841.

Case, Karl and Robert Shiller (Nov. 1987). *Prices of Single Family Homes Since 1970: New Indexes for Four Cities*. Working Paper 2393. National Bureau of Economic Research. DOI: 10.3386/w2393.

— (Feb. 1988). *The Efficiency of the Market for Single-Family Homes*. Working Paper 2506. National Bureau of Economic Research. DOI: 10.3386/w2506.

Cellmer, Radosław, Aneta Cichulska, and Mirosław Bełej (2020). "Spatial Analysis of Housing Prices and Market Activity with the Geographically Weighted Regression". In: *ISPRS International Journal of Geo-Information* 9.6. DOI: 10.3390/ijgi9060380.

Chandler, Daniel and Richard Disney (2014). *Measuring house prices: a comparison of different indices*. Institute for Fiscal Studies. ISBN: 978-1-909463-39-4. URL: https://ifs.org.uk/publications/measuring-house-prices-comparison-different-indices.

Che, Xinwei, Bin Li, et al. (2011). "Property Prices and Bank Lending: Some Evidence from China's Regional Financial Centres". In: *Proceedings of the International Conference on Computational Science* 4, pp. 1660–1667. DOI: 10.1016/j.procs.2011.04.179.

Chen, Ming-Chi, Yuichiro Kawaguchi, and Kanak Patel (2004). "An analysis of the trends and cyclical behaviours of house prices in the Asian markets". In: *Journal of Property Investment & Finance* 22.1, pp. 55–75. DOI: 10.1108/14635780410525144.

Choueifaty, Yves, Tristan Froidure, and Julien Reynier (2013). "Properties of the most diversified portfolio". In: *Journal of Investment Strategies* 2.2, pp. 49–70. DOI: 10.2139/ssrn.1895459.

Clapp, John M and Carmelo Giaccotto (1992). "Estimating price trends for residential property: a comparison of repeat sales and assessed value methods". In: *The Journal of Real Estate Finance and Economics* 5.4, pp. 357–374. DOI: 10.1007/BF00174805.

Clapp, John M., Hyon–Jung Kim, and Alan E. Gelfand (2002). "Predicting Spatial Patterns of House Prices Using LPR and Bayesian Smoothing". In: *Real Estate Economics* 30.4, pp. 505–532. DOI: 10.1111/1540-6229.00048.

Conefrey, Thomas, David Staunton, et al. (2019). "Population change and housing demand in Ireland". In: *Central Bank of Ireland Economic Letter* 14, pp. 1–16.

Conway, Delores, Christina Q Li, et al. (2010). "A spatial autocorrelation approach for examining the effects of urban greenspace on residential property values". In: *The Journal of Real Estate Finance and Economics* 41, pp. 150–169. DOI: 10.1007/s11146-008-9159-6.

Costello, Greg, Patricia Fraser, and Nicolaas Groenewold (2011). "House prices, non-fundamental components and interstate spillovers: The Australian experience". In: *Journal of Banking and Finance* 35.3, pp. 653–669. DOI: 10.1016/j.jbankfin.2010.07.035.

Costello, Greg and Craig Watkins (2002). "Towards a System of Local House Price Indices". In: *Housing Studies* 17.6, pp. 857–873. DOI: 10.1080/02673030216001.

Dalton, Padraig and Ken Moore (Dec. 2014). *How to quickly adapt to new policy needs? The experience of the Central Statistics Office, Ireland in developing house price indicators*. URL: https://www.ine.pt/scripts/DGINS-2015/presentations/S1_P3_CSO.pdf.

De La Briandais, Rene (1959). "File Searching Using Variable Length Keys". In: IRE-AIEE-ACM '59 (Western). Association for Computing Machinery, 295–298. ISBN: 9781450378659. DOI: 10.1145/1457838.1457895.

De Vries, Paul, Jan de Haan, et al. (2009). "A house price index based on the SPAR method". In: *Journal of Housing Economics* 18.3. Special Issue on Owner Occupied Housing in National Accounts and Inflation Measures, pp. 214–223. ISSN: 1051-1377. DOI: 10.1016/j.jhe.2009.07.002.

De Wit, Erik R., Peter Englund, and Marc K. Francke (2013). "Price and transaction volume in the Dutch housing market". In: *Regional Science and Urban Economics* 43.2, pp. 220–241. DOI: 10.1016/j.regsciurbeco.2012.07.002.

Diewert, W Erwin and Kevin J Fox (2022). "Substitution bias in multilateral methods for cpi construction". In: *Journal of Business & Economic Statistics* 40.1, pp. 355–369. DOI: 10.1080/07350015.2020.1816176.

Ding, Jiale, Wenying Cen, et al. (2024). "A neural network model to optimize the measure of spatial proximity in geographically weighted regression approach: a case study on house price in Wuhan". In: *International Journal of Geographical Information Science* 38.7, pp. 1315–1335. DOI: 10.1080/13658816.2024.2343771.

Domingos, Pedro (2012). "A few useful things to know about machine learning". In: *Communications of the ACM* 55.10, pp. 78–87. DOI: 10.1145/2347736.2347755.

Dorsey, Robert E., Haixin Hu, et al. (2010). "Hedonic versus repeat-sales housing price indexes for measuring the recent boom-bust cycle". In: *Journal of Housing Economics* 19.2, pp. 75–93. ISSN: 1051-1377. DOI: 10.1016/j.jhe.2010.04.001.

Elul, Ronel, Nicholas S Souleles, et al. (2010). "What "triggers" mortgage default?" In: *American Economic Review* 100.2, pp. 490–94. DOI: 10.1257/aer.100.2.490.

Englund, Peter, Min Hwang, and John M. Quigley (Jan. 2002). "Hedging Housing Risk". In: *The Journal of Real Estate Finance and Economics* 24.1, pp. 167–200. ISSN: 1573-045X. DOI: 10.1023/A:1013942607458.

Englund, Peter and Yannis M Ioannides (1997). "House price dynamics: an international empirical perspective". In: *Journal of Housing Economics* 6.2, pp. 119–136. DOI: 10.1006/jhec.1997.0210.

Falzon, Joseph and David Lanzon (2013). "Comparing alternative house price indices: evidence from asking prices in Malta". In: *International Journal of Housing Markets and Analysis*. DOI: 10.1108/17538271311306048.

Gillen, Kevin, Thomas Thibodeau, and Susan Wachter (2001). "Anisotropic autocorrelation in house prices". In: *The Journal of Real Estate Finance and Economics* 23, pp. 5–30. DOI: 10.1023/A:1011140022948.

Goh, Yen Min, Greg Costello, and Greg Schwann (2012). "Accuracy and Robustness of House Price Index Methods". In: *Housing Studies* 27.5, pp. 643–666. DOI: 10.1080/02673037.2012.697551.

Goldberg, Stephen R, Mary J Phillips, and H James Williams (2009). "Survive the recession by managing cash". In: *Journal of Corporate Accounting & Finance* 21.1, pp. 3–9. DOI: 10.1002/jcaf.20540.

Gouriéroux, Christian and Anne Laferrère (2009). "Managing hedonic housing price indexes: The French experience". In: *Journal of Housing Economics* 18.3, pp. 206–213. DOI: 10.1016/j.jhe.2009.07.012.

Grover, Richard and Chris Grover (Aug. 2013). "Property cycles". In: *Journal of Property Investment and Finance* 31.5, pp. 502–516. DOI: 10.1108/JPIF-05-2013-0030.

Guerrieri, Veronica and Harald Uhlig (2016). "Housing and Credit Markets: Booms and Busts". In: vol. 2. Handbook of Macroeconomics. Elsevier, pp. 1427–1496. DOI: 10.1016/bs.hesmac.2016.06.001.

Guo, Yi, Stephen Tierney, and Junbin Gao (2021). "Efficient sparse subspace clustering by nearest neighbour filtering". In: *Signal Processing* 185, p. 108082. ISSN: 0165-1684. DOI: 10.1016/j.sigpro.2021.108082.

Guttman, Antonin (June 1984). "R-trees: A Dynamic Index Structure for Spatial Searching". In: *SIGMOD Rec.* 14.2, pp. 47–57. ISSN: 0163-5808. DOI: 10.1145/971697.602266.

Haan, Jan de and WE Diewert (2011). "Handbook on residential property price indexes". In: *Luxembourg: Eurostat*.

Hand, David J. (2013). "Data Mining". In: *Encyclopedia of Environmetrics*. American Cancer Society. ISBN: 9780470057339. DOI: 10.1002/9780470057339.vad002.pub2.

Hansen, James (2009). "Australian house prices: a comparison of hedonic and repeat-sales measures". In: *Economic Record* 85.269, pp. 132–145. DOI: 10.1111/j.1475-4932.2009.00544.x.

Hatzvi, Eden and Glenn Otto (2008). "Prices, rents and rational speculative bubbles in the Sydney housing market". In: *Economic Record* 84.267, pp. 405–420. DOI: 10.1111/j.1475-4932.2008.00484.x.

Haurin, Donald R. and David Brasington (1996). "School Quality and Real House Prices: Inter- and Intrametropolitan Effects". In: *Journal of Housing Economics* 5.4, pp. 351–368. DOI: 10.1006/jhec.1996.0018.

He, Chengjie, Zhen Wang, et al. (2010). "Driving Forces Analysis for Residential Housing Price in Beijing". In: *Procedia Environmental Sciences* 2. International Conference on Ecological Informatics and Ecosystem Conservation (ISEIS 2010), pp. 925–936. ISSN: 1878-0296. DOI: 10.1016/j.proenv.2010.10.104.

Henneberry, John (1998). "Transport investment and house prices". In: *Journal of property valuation and investment* 16.2, pp. 144–158. DOI: 10.1108/14635789810212913.

Hess, Andreas and Arne Holzhausen (2008). *The structure of European mortgage markets*. Tech. rep. Working paper.

Hill, Robert, Michael Scholz, et al. (Oct. 2018). "An evaluation of the methods used by European countries to compute their official house price indices". In: *Economie et Statistique / Economics and Statistics* 2018, pp. 221–238. DOI: 10.24187/ecostat.2018.500t.1953.

Hill, Robert J, Norbert Pfeifer, et al. (2024). "Warning: Some transaction prices can be detrimental to your house price index". In: *Review of Income and Wealth* 70.2, pp. 320–344.

HM Land Registry (2024). *Price Paid dataset*. Accessed: 2024-06-01. URL: https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads.

Iacoviello, Matteo and Francois Ortalo-Magne (2003). "Hedging housing risk in London". In: *The Journal of Real Estate Finance and Economics* 27, pp. 191–209. DOI: 10.1023/A:1024776303998.

II, John A. Pearce and Steven C. Michael (2006). "Strategies to prevent economic recessions from causing business failure". In: *Business Horizons* 49.3, pp. 201–209. ISSN: 0007-6813. DOI: 10.1016/j.bushor.2005.08.008.

Ismail, Suriatini (Jan. 2006). "Spatial autocorrelation and real estate studies: A literature review". In: *Regional Science and Urban Economics* 35.

Jansen, Sylvia, P Vries, et al. (Jan. 2008). "Developing a House Price Index for The Netherlands: A Practical Application of Weighted Repeat Sales". In: *The Journal of Real Estate Finance and Economics* 37, pp. 163–186. DOI: 10.1007/s11146-007-9068-0.

Jones, Colin, Stewart Cowe, and Edward Trevillion (2018). *Property boom and banking bust: The role of commercial lending in the bankruptcy of banks*. John Wiley & Sons. ISBN: 978-1-119-21925-5.

Jones, Phil and James Evans (2013). *Urban regeneration in the UK: Boom, bust and recovery*. Sage. ISBN: 978-1-473-91501-5. DOI: 10.4135/9781473915015.

Kain, John F. and John M. Quigley (1970). "Measuring the Value of Housing Quality". In: *Journal of the American Statistical Association* 65.330, pp. 532–548. DOI: 10.1080/01621459.1970.10481102.

Kennedy, Gerard and Samantha Myers (Nov. 2019). *An overview of the Irish housing market*. Financial Stability Notes 16/FS/19. Central Bank of Ireland. URL: https://www.centralbank.ie/docs/default-source/publications/financial-stability-notes/no-16-an-overview-of-the-irish-housing-market-(kennedy-and-myers).pdf.

Kothuri, Ravi Kanth V, Siva Ravada, and Daniel Abugov (2002). "Quadtree and R-tree indexes in oracle spatial: a comparison using GIS data". In: *2002 ACM SIGMOD International Conference on Management of Data*. ACM, pp. 546–557. DOI: 10.1145/564691.564755.

Kuo, Li-Lan and Feifei Li (2013). "An Investor's Low Volatility Strategy". In: *The Journal of Index Investing* 3.4, pp. 8–22. DOI: 10.3905/jii.2013.3.4.008.

Labonte, Marc (2007). *Would a Housing Crash Cause a Recession?* Congressional Research Service.

Larson, William D. and Justin Contat (Apr. 2021). *Transaction Composition and House Price Index Measurement: Evidence from a Repeat-Sales Aggregation Index*. FHFA Staff Working Papers 21-01. Federal Housing Finance Agency. URL: https://ideas.repec.org/p/hfa/wpaper/21-01.html.

Law, Stephen (2017). "Defining Street-based Local Area and measuring its effect on house price using a hedonic price approach: The case study of Metropolitan London". In: *Cities* 60, pp. 166–179. ISSN: 0264-2751. DOI: 10.1016/j.cities.2016.08.008.

Lee, Jae-Gil and Minseo Kang (2015). "Geospatial Big Data: Challenges and Opportunities". In: *Big Data Research* 2.2. Visions on Big Data, pp. 74–81. ISSN: 2214-5796. DOI: 10.1016/j.bdr.2015.01.003.

Leung, Tin Cheuk and Kwok Ping Tsang (2013). "Anchoring and loss aversion in the housing market: Implications on price dynamics". In: *China Economic Review* 24, pp. 42–54. DOI: 10.1016/j.chieco.2012.10.003.

Liow, Kim and Joseph Ooi (Oct. 2004). "Does corporate real estate create wealth for shareholders?" In: *Journal of Property Investment and Finance* 22, pp. 386–400. DOI: 10.1108/14635780410556870.

Luttik, Joke (2000). "The value of trees, water and open space as reflected by house prices in the Netherlands". In: *Landscape and Urban Planning* 48.3, pp. 161–167. DOI: 10.1016/S0169-2046(00)00039-6.

Maguire, Phil, Stephen Kelly, et al. (2017). "Further evidence in support of a low-volatility anomaly: Optimizing buy-and-hold portfolios by minimizing historical aggregate volatility". In: *Journal of Asset Management* 18, pp. 326–339. DOI: 10.1057/s41260-016-0036-1.

Maguire, Phil, Robert Miller, et al. (2016). "A robust house price index using sparse and frugal data". In: *Journal of Property Research* 33.4, pp. 293–308. DOI: 10.1080/09599916.2016.1258718.

Maguire, Phil, Philippe Moser, et al. (2014). "Maximizing positive porfolio diversification". In: *2014 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*. IEEE, pp. 174–181. DOI: 10.1109/CIFEr.2014.6924070.

Mar Iman, Abdul Hamid (2001). "Incorporating a geographic information system in the hedonic modelling of farm property values". PhD thesis. Lincoln University. URL: https://hdl.handle.net/10182/2161.

Maslow, A. H. (July 1943). "A theory of human motivation." In: *Psychological Review* 50.4, pp. 370–396. DOI: 10.1037/h0054346.

McDonald, Chris, Mark Smith, et al. (2009). *Developing stratified housing price measures for New Zealand*. Tech. rep. Reserve Bank of New Zealand. URL: https://www.rbnz.govt.nz/hub/publications/discussion-paper/2009/dp2009-07.

McGinnis, Will (2017). *Pygeohash*. [Python]. URL: https://github.com/wdm0006/pygeohash.

McKee, Kim (2012). "Young People, Homeownership and Future Welfare". In: *Housing Studies* 27.6, pp. 853–862. DOI: 10.1080/02673037.2012.714463.

McMillen, Daniel P. (Jan. 2003). "Neighborhood house price indexes in Chicago: a Fourier repeat sales approach". In: *Journal of Economic Geography* 3.1, pp. 57–73. ISSN: 1468-2702. DOI: 10.1093/jeg/3.1.57.

McMillen, Daniel P. and Paul Thorsnes (2006). "Housing Renovations and the Quantile Repeat-Sales Price Index". In: *Real Estate Economics* 34.4, pp. 567–584. DOI: 10.1111/j.1540-6229.2006.00179.x.

Miller, Robert and Phil Maguire (2020). "A rapidly updating stratified mix-adjusted median property price index model". In: *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, pp. 9–15. DOI: 10.1109/SSCI47803.2020.9308235.

— (2021). "GeoTree: A Data Structure for Constant Time Geospatial Search Enabling a Real-Time Property Index". In: *Intelligent Computing: Proceedings of the 2021 Computing Conference, Volume 2*. Springer, pp. 152–165. DOI: 10.1007/978-3-030-80126-7_12.

— (2022). "A real-time mix-adjusted median property price index enabled by an efficient nearest neighbour approximation data structure". In: *Journal of Banking and Financial Technology*, pp. 1–14. DOI: 10.1007/s42786-022-00043-y.

Moussalli, R., M. Srivatsa, and S. Asaad (May 2015). "Fast and Flexible Conversion of Geohash Codes to and from Latitude/Longitude Coordinates". In: *2015 IEEE 23rd Annual International Symposium on Field-Programmable Custom Computing Machines*, pp. 179–186. DOI: 10.1109/FCCM.2015.18.

Moussalli, Roger, Sameh W. Asaad, and Mudhakar Srivatsa (2015). "Enhanced conversion between geohash codes and corresponding longitude/latitude coordinates". Pat. US20160283515A1. URL: https://patents.google.com/patent/US20160283515.

Muja, M. and D. G. Lowe (Nov. 2014). "Scalable Nearest Neighbor Algorithms for High Dimensional Data". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.11, pp. 2227–2240. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2014.2321376.

MyHome Ltd. (2024). Accessed: 2024-06-01. URL: http://www.myhome.ie.

Nationwide (Jan. 2024). *House price index methodology*. URL: https://www.nationwide.co.uk/-/assets/nationwidecouk/documents/about/house-price-index/nationwide-hpi-methodology.pdf.

Niemeyer, Gustavo (2008). *geohash.org is public!* Accessed: 2024-06-01. URL: https://blog.labix.org/2008/02/26/geohashorg-is-public.

Nofsinger, John R. (2012). "Household behavior and boom/bust cycles". In: *Journal of Financial Stability* 8.3. The Financial Crisis of 2008, Credit Markets and Effects on Developed and Emerging Economies, pp. 161–173. ISSN: 1572-3089. DOI: 10.1016/j.jfs.2011.05.004.

Norman Mille Vivek Sah, Michael Sklarz and Stefan Pampulov (2013). "Is there Seasonality in Home Prices—Evidence from CBSAs". In: *Journal of Housing Research* 22.1, pp. 1–15. DOI: 10.1080/10835547.2013.12092066.

OECD, Eurostat, et al. (2013). *Handbook on Residential Property Price Indices*, p. 186. DOI: 10.1787/9789264197183-en.

O'Hanlon, Niall (2011). "Constructing a national house price index for Ireland". In: *Journal of the Statistical and Social Inquiry Society of Ireland* 40, pp. 167–196. ISSN: 00814776. URL: http://hdl.handle.net/2262/62349.

ONS (Dec. 2023a). *UK House Price Index: Acorn (CACI) Consumer Classification Study*. Tech. rep. URL: https://www.gov.uk/government/statistics/quality-assurance-of-administrative-data-in-the-uk-house-price-index/acorn-consumer-classification-caci.

— (Dec. 2023b). *UK House Price Index: Quality and methodology*. Tech. rep. URL: https://www.gov.uk/government/publications/about-the-uk-house-price-index/quality-and-methodology.

— (Dec. 2023c). *UK House Price Index: Scottish Energy Performance Certificate Study*. Tech. rep. URL: https://www.gov.uk/government/statistics/quality-assurance-of-administrative-data-in-the-uk-house-price-index/scottish-energy-performance-certificates.

— (Dec. 2023d). *UK House Price Index: Valuation Office Agency Council Tax Valuation List Study*. Tech. rep. URL: https://www.gov.uk/government/statistics/quality-assurance-of-administrative-data-in-the-uk-house-price-index/valuation-office-agency-council-tax-valuation-lists.

Ortalo-Magné, François and Sven Rady (Apr. 2006). "Housing Market Dynamics: On the Contribution of Income Shocks and Credit Constraints*". In: *The Review of Economic Studies* 73.2, pp. 459–485. ISSN: 0034-6527. DOI: 10.1111/j.1467-937X.2006.383_1.x.

Paci, Lucia, María Asunción Beamonte, et al. (2017). "Analysis of residential property sales using space–time point patterns". In: *Spatial Statistics* 21, pp. 149 –165. ISSN: 2211-6753. DOI: 10.1016/j.spasta.2017.06.007.

Piddington, Justine, Simon Nicol, et al. (2020). *The Housing Stock of the United Kingdom*. Tech. rep. BRE Trust, UK. URL: https://files.bregroup.com/bretrust/The-Housing-Stock-of-the-United-Kingdom_Report_BRE-Trust.pdf.

Prasad, Nalini and Anthony Richards (2008). "Improving Median Housing Price Indexes through Stratification". In: *Journal of Real Estate Research* 30.1, pp. 45–72. DOI: 10.1080/10835547.2008.12091213.

Property Services Regulatory Authority (IE) (June 2024). *Residential Property Price Register*. URL: https://www.propertypriceregister.ie/website/npsra/pprweb.nsf/PPR?OpenForm.

Quigley, John M. (1995). "A Simple Hybrid Model for Estimating Real Estate Price Indexes". In: *Journal of Housing Economics* 4.1, pp. 1–12. ISSN: 1051-1377. DOI: 10.1006/jhec.1995.1001.

— (1999). "Real Estate Prices and Economic Cycles". In: *International Real Estate Review* 2.1, pp. 1–20. URL: https://ideas.repec.org/a/ire/issued/v02n011999p1-20.html.

Ramírez-Gallego, Sergio, Bartosz Krawczyk, et al. (2017). "Nearest Neighbor Classification for High-Speed Big Data Streams Using Spark". In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 47.10, pp. 2727–2739. DOI: 10.1109/TSMC.2017.2700889.

Rappaport, Jordan et al. (2007). "A guide to aggregate house price measures". In: *Economic Review - Federal Reserve Bank of Kansas City* 92.2, pp. 41–71. URL: https://www.kansascityfed.org/documents/1400/2007-A%20Guide%20to%20Aggregate%20House%20Price%20Measures.pdf.

Reusens, Peter, Frank Vastmans, and Sven Damen (2023). "A new framework to disentangle the impact of changes in dwelling characteristics on house price indices". In: *Economic Modelling* 123, p. 106252. DOI: 10.1016/j.econmod.2023.106252.

Robusto, C. C. (1957). "The Cosine-Haversine Formula". In: *The American Mathematical Monthly* 64.1, pp. 38–40. DOI: 10.2307/2309088.

Roussopoulos, Nick, Stephen Kelley, and Frédéric Vincent (May 1995). "Nearest Neighbor Queries". In: *SIGMOD Rec.* 24.2, pp. 71–79. ISSN: 0163-5808. DOI: 10.1145/568271.223794.

Rymon, Ron (1992). "Search through systematic set enumeration". In: *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning*. KR'92. Morgan Kaufmann Publishers Inc., 539–550. ISBN: 1558602623.

Safar, Maytham (Aug. 2005). "K nearest neighbor search in navigation systems". In: *Mob. Inf. Syst.* 1.3, 207–224. ISSN: 1574-017X. DOI: 10.1155/2005/692568.

Sayag, Doron, Dano Ben-hur, and Danny Pfeffermann (2022). "Reducing revisions in hedonic house price indices by the use of nowcasts". In: *International Journal of Forecasting* 38.1, pp. 253–266. ISSN: 0169-2070. DOI: 10.1016/j.ijforecast.2021.04.008.

Scatigna, Michela, Robert Szemere, and Kostas Tsatsaronis (Sept. 2014). *Residential property price statistics across the globe*. Quarterly Review. Bank for International Settlements. URL: https://www.bis.org/publ/qtrpdf/r_qt1409h.htm.

Shiller, Robert J. (Mar. 2003). "From Efficient Markets Theory to Behavioral Finance". In: *Journal of Economic Perspectives* 17.1, 83–104. DOI: 10.1257/089533003321164967.

Silver, Mick (Nov. 2016). *How to better measure hedonic residential property price indexes*. Working Paper series WP/16/213. International Monetary Fund. URL: https://www.imf.org/external/pubs/ft/wp/2016/wp16213.pdf.

Soltani, Ali, Nader Zali, et al. (2023). "Multilevel impacts of urban amenities on housing price in Tehran, Iran". In: *Journal of Urban Planning and Development* 149.4, p. 05023028. DOI: 10.1061/JUPDDM.UPENG-4434.

Sommervoll, Dag Einar (2006). "Temporal aggregation in repeated sales models". In: *The Journal of Real Estate Finance and Economics* 33, pp. 151–165. DOI: 10.1007/s11146-006-8946-1.

Tu, Yong, Seow Eng Ong, and Ying Hua Han (2009). "Turnovers and housing price dynamics: Evidence from Singapore condominium market". In: *The Journal of Real Estate Finance and Economics* 38.3, pp. 254–274. DOI: 10.1007/s11146-008-9155-x.

Turner, Anthony G (2003). "Sampling strategies". In: *Handbook on designing of household sample surveys*.

Verbic, Miroslav and Peter Korenčan (June 2017). "Cluster-based econometric analysis of house prices in Slovenia". In: *Geodetski Vestnik* 61, pp. 231–245. DOI: 10.15292/geodetski-vestnik.2017.02.231-245.

Wallace, Nancy E and Richard A Meese (1997). "The construction of residential housing price indices: a comparison of repeat-sales, hedonic-regression, and hybrid approaches". In: *The Journal of Real Estate Finance and Economics* 14, pp. 51–73. DOI: 10.1023/A:1007715917198.

Wang, Ferdinand T. and Peter M. Zorn (1997). "Estimating House Price Growth with Repeat Sales Data: What's the Aim of the Game?" In: *Journal of Housing Economics* 6.2, pp. 93–118. ISSN: 1051-1377. DOI: 10.1006/jhec.1997.0209.

Wood, Robert (Dec. 2005). "A comparison of UK residential house price indices". In: *Real estate indicators and financial stability*. Ed. by Bank for International Settlements. Vol. 21. BIS Papers chapters. Bank for International Settlements, pp. 212–227. URL: https://www.bis.org/publ/bppdf/bispap21p.pdf.

Zhang, Jun, N. Mamoulis, et al. (2004). "All-nearest-neighbors queries in spatial databases". In: *Proceedings. 16th International Conference on Scientific and Statistical Database Management, 2004.* Pp. 297–306. DOI: 10.1109/SSDM.2004.1311221.

Zhu, Haibin (Dec. 2005). "The importance of property markets for monetary policy and financial stability". In: *Real estate indicators and financial stability*. Ed. by Bank for International Settlements. Vol. 21. BIS Papers chapters. Bank for International Settlements, pp. 9–29. URL: https://www.bis.org/publ/bppdf/bispap21c.pdf.