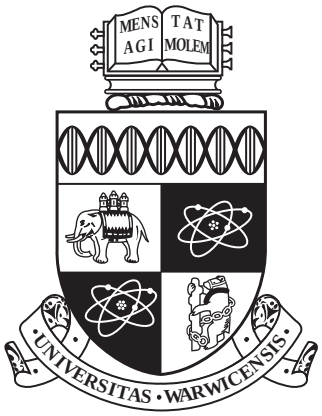


Road Distance and Travel Time for Spatial Urban Modelling

by

Henry Crosby

A thesis submitted to The University of Warwick

in partial fulfilment of the requirements

for admission to the degree of

Doctor of Philosophy

Department of Computer Science

The University of Warwick

September 2018

Abstract

Interactions within and between urban environments include the price of houses, the flow of traffic and the intensity of noise pollution, which can all be restricted by various physical, regulatory and customary barriers. Examples of such restrictions include buildings, one-way systems and pedestrian crossings. These constrictive features create challenges for predictive modelling in urban space, which are not fully captured when proximity-based models rely on the typically used Euclidean (straight line) distance metric.

Over the course of this thesis, I ask three key questions in an attempt to identify how to improve spatial models in restricted urban areas. These are: (1) which distance function best models real world spatial interactions in an urban setting? (2) when, if ever, are non-Euclidean distance functions valid for urban spatial models? and (3) what is the best way to estimate the generalisation performance of urban models utilising spatial data?

This thesis answers each of these questions through three contributions supporting the interdisciplinary domain of Urban Sciences. These contributions are: (1) the provision of an improved approximation of road distance and travel time networks to model urban spatial interactions; (2) the approximation of valid distance metrics from non-Euclidean inputs for improved spatial predictions and (3) the presentation of a road distance and travel time cross-validation metric to improve the estimation of urban model generalisation. Each of these contributions provide improvements against the current state-of-the-art. Throughout, all experiments utilise real world datasets in England and Wales, such datasets contain information on restricted roads, travel times, house sales and traffic counts. With these datasets, I display a number of case studies which show up to a 32% improved model accuracy against Euclidean distances and in some cases, a 90% improvement for the estimation of model generalisation performance.

Combined, the contributions improve the way that proximity-based urban models perform and also provides a more accurate estimate of generalisation performance for predictive models in urban space. The main implication of these contributions to Urban Science is the ability to better model the challenges within a city based on how they interact with themselves and each other using an improved function of urban mobility, compared with the current state-of-the-art. Such challenges may include selecting the optimal locations for emergency services, identifying the causes of traffic incidents or estimating the density of air pollution. Additionally, the key implication of this research on geostatistics is that it provides the motivation and means of undertaking non-Euclidean

based research for non-urban applications, for example predicting with alternative, non-road based, mobility patterns such as migrating animals, rivers and coast lines. Finally, the implication of my research to the real estate industry is significant, in which one can now improve the accuracy of the industry's state-of-the-art nationwide house price predictor, whilst also being able to more appropriately present their accuracy estimates for robustness.

Dedicated to...
everyone that has ever believed in me.

Acknowledgements

My special thanks go to **Dr. Theodoros Damoulas** who has supported my technical and written work far beyond the call of duty. You have enabled me not only to complete my doctoral thesis, but to challenge it, which in turn has shown me to see the true joy in research. Another huge thank you also goes to **Professor João Porto de Albuquerque** who has opened up my world of research to new areas and supported me through all my different endeavours. My academic career so far would be non-existent without **Professor Stephen Jarvis** who set me on my path of research and introduced me to some fantastic opportunities for personal growth and applications.

In addition, I would like to thank **Professor Celia Lury** and **Dr. Narushige Shiode** who each in turn saw my potential at the early stages, allowing me to enter a university department and research area which may not have otherwise been accessible to me. Furthermore, the support of **Yvonne Colmer** and **Katie Martin** is truly appreciated. Additionally, **the Warwick Urban Science Doctoral Programme** provides a supportive community of like minded researchers - thank you to all within it.

I would like to express my special thanks to **Elizabeth**, **Chris**, **Libby**, **Kev** and **Sarah** for all of the good times and your loving help throughout. My dearest **Claire**, you have been the most amazing wife I could ever have asked for over the past three years. Not only have you supported me but you have made my every day interesting, exciting and wonderful - here is to a lifetime more of these days! Finally, my parents **Jane** and **Rob** - you have taught, supported and believed in my abilities every day and for that, this thesis is dedicated you.

My sincerest gratitude goes to you all!

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. Parts of this thesis have been published by the author:

1. Crosby, H., Davis, P. and Jarvis, S.A., 2015, November. Exploring New Data Sources to Improve UK Land Parcel Valuation. In Proceedings of the 1st International ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics (pp. 32-35). ACM.

2. Crosby, H., Davis, P. and Jarvis, S.A., 2016, September. Spatially-Intensive Decision Tree Prediction of Traffic Flow across the entire UK Road Network. IEEE/ACM 20th International Symposium on International Symposium on Distributed Simulation and Real Time Applications (pp. 116-119).

3. Crosby, H., Davis, P., Damoulas, T. and Jarvis, S.A., 2016, October. A spatio-temporal, Gaussian process regression, real-estate price predictor. In Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (p. 68). ACM.

4. Crosby, H., Damoulas, T., Porto de Albuquerque J., Caton, A., and Jarvis, S.A., 2018, May. Road distance and travel time for an improved house price Kriging predictor, Geo-spatial Information Science, 21:3, 185-194, DOI: 10.1080/10095020.2018.1503775

5. Crosby H., Damoulas T and Jarvis S.A., 2019. Embedding road networks and travel time into distance metrics for urban modelling, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2018.1547386

6. Crosby, H., Damoulas, T. and Jarvis, S.A., 2018, November. Road and Travel Time Validation for Urban Modelling. Final draft completed for submission to the International Journal of Geographic Information Sciences (IJGIS).

Sponsorship and Grants

The research presented in this thesis was made possible by the support of the following benefactors and sources:

- Sponsor 1 : Engineering and Physical Sciences Research Council (EPSRC) Centre for Doctoral Training in Urban Science (EP/ L016400/ 1). 2014-2018.
- Sponsor 2 : Assured Property Group, Innovation Center, Warwick Innovation Center, Gallows Hill, Warwick, CV34 6UW. 2015-2016.

Abbreviations

i.i.d independent and identically distributed

UK United Kingdom

AI artificial intelligence

ML machine learning

OSM OpenStreetMaps

OSRM Open Street Routing Machine

SAC spatial autocorrelation

RQ research questions

PD positive definite

PSD positive semi-definite

CND conditionally negative definite

ROI return on investment

CPT Central Place Theory

AVM automated valuation model

AVMs automated valuation models

ITS intelligent traffic systems

GIS Geographic Information Systems

long longitude

lat latitude

RMSE Root Mean Squared Error

NRMSE Normalised Root Mean Squared Error

MAPE Mean Absolute Percentage Error

r^2 the squared Pearson correlation coefficient

GWR geographically weighted regression

OS's Ordnance Survey's

OS Ordnance Survey

FW Floyd Warshall

MDS multidimensional scaling

HMLR's Her Majesty's Land Registry's

WGS World Geodetic System

KCV k-fold cross validation

S-KCV spatial k-fold cross validation

RT-KCV road distance and travel time k-fold cross validation

CV cross validation

GPR Gaussian process regression

Symbols

Ω	sample space of possible outcomes
F	set of possible events
P	probability measure over Ω
$Z(s)$	random variable at location s
s	spatial data point
D	full spatial dataset
T	full temporal dataset
$\mathbb{R}$	set of real numbers
$\mathbb{Q}$	set of rational numbers
$\mathbb{N}$	set of natural numbers
$\bar{s}$	average of all points s
γ	semivariance
$\mathbf{s}$	vector of points s
$d_{i,j}$	distance between points s_i and s_j
r^2	goodness of fit between two point sets
ρ	deadzone radius
RD	road distance
TT	travel time
ζ	metric space
λ_i	eigenvalue i
δ	approximate road distance
τ	approximate travel time

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

By 2030, it is expected that 5 billion people will live in urban spaces, 662 cities will have at least 1 million residents and there will be a total urban spread of 1.2 million km^2 [12, 98, 114]. Hence, cities will continue to accommodate over 50% of the world's population. In the United Kingdom (UK) however, the proportional population is hugely extenuated with over 82% of UK citizens already living in urban spaces. This population resides in 64 cities and has grown by more than 13% in the past 30 years [18]. Many UK cities suffer from legacy infrastructure - the City of London, for example, relies on sewage infrastructure originally built in the 1860s - which impacts on their ability to support projected growth. Such challenges are well documented: housing supply is not matching demand [62]; commuting times are increasing [54] and there are shortages in services for the most vulnerable citizens [116]. It is indeed these issues that reflect the very nature of the growing UK city and the motivation to undertake analysis for urban sustainability [124].

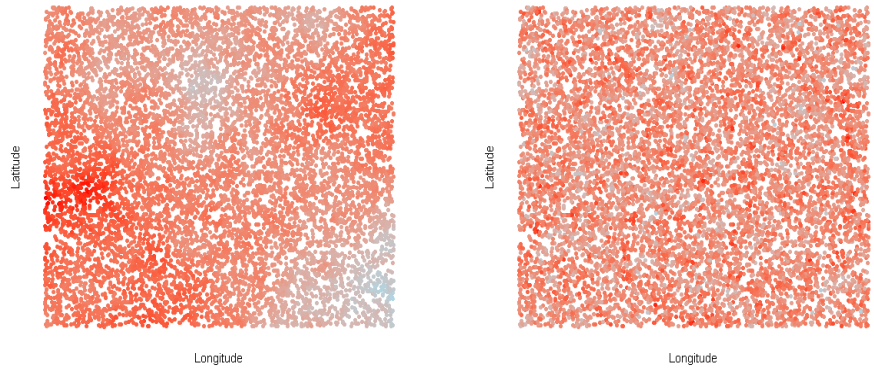
Annually, digital urban data witnesses a 42% compound growth [52] and produces more information about the social and physical structures of contemporary cities than have ever before been available. In fact, before the new decade, urban systems were inferred by (relatively) small scale data, which were sourced from traditional data collection methods such as surveys, questionnaires, interviews or observations. Although detailed, these methods could sometimes lack the representativeness and reliability that 'big data' boasts. This aforementioned 'big data' is typically (1) crowdsourced voluntarily, such as OpenStreetMaps (OSM), Wikipedia and Youtube; (2) outsourced to citizens (albeit, sometimes unknowingly to the user) including mobile data, security

footage, road sensors, social networks and on-line shopping or (3) sourced from the government, for example house sale prices, administrative regions, air pollution and land use. Each of these geotagged and/or timestamped data observe the various aspects of the lives of urban citizens at a higher spatial and temporal resolution than ever before.

Commonly, the flow within cities is referred to as its metabolism [5] and the larger a city gets the more interrelated and diverse that metabolism becomes. A large and complex metabolism of diversity, networks and citizens can hence become a power in itself [105] which must be managed [9] to be understood and to ensure that our cities of the future are sustainable, efficient and promoting of a high quality of life [119]. Consequently, the opportunity that the aforementioned urban data can provide is the potential to rationalise and simplify the otherwise complex and multivariate processes that form a city, for example road networks, cultural diversity and citizen wealth [128]. In other words, the data facilitates the means to answering the question 'how do cities work?'. The interdisciplinary area of research interested in addressing this question is *Urban Science*, a discipline heavily motivated by data and quantitative urban models.

Contemporary Urban Science is primarily concerned with modelling urban spaces, of which the purpose is two fold; explanatory and predictive. Urban models are holistic and unique in the fact that they utilise state-of-the-art technologies (i.e., machine learning) from the viewpoint of a geographer, policy maker or economist for example. Modelling cities without these traditional discipline boundaries helps to maintain the sustainable growth of urban-specific innovations by domain rather than discipline. Specifically, urban models attempt to obtain the relationship between some target value, for instance the price of houses, and some other variable(s) such as topography [73], building footprints [102] or crime [121]. Interestingly, space [37] followed by time [65] consistently define the largest proportions of most urban models, for example house prices [36], traffic flow [136] and well-being [64].

Spatial urban models typically perform well given that geographical proxim-

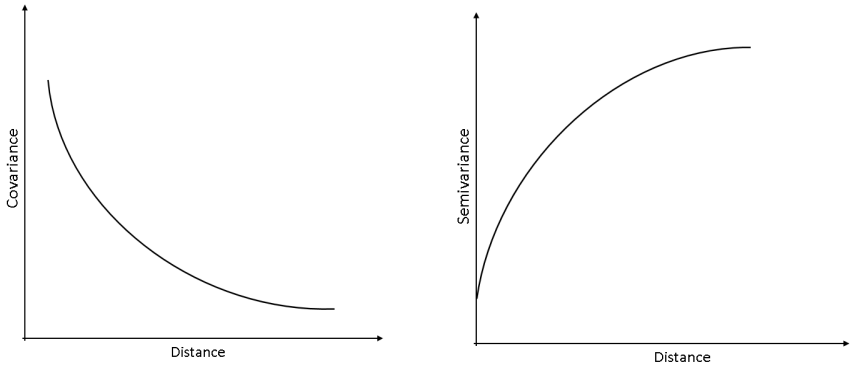


(a) SAC present (b) No SAC present.

Figure 1.1: A visual representation of SAC. Figure (a) shows a strong SAC and Figure (b) shows a weak SAC.

ity intuitively measures urban processes which are defined as the set of interactions that measure the patterns of flow and networks of relations in a city [10]. As such, observations in our inherently spatial cities can be probabilistically determined by SAC: the similarity between two observations as a function of geographical proximity. Figure 1.1 is a visualisation of this concept where each point represents a spatial location for some simulated observation and each colour represents a value for each observation from low (blue) to high (red). Such spatial relationships violate the typical assumption present in non-spatial statistics; all observations are independent and identically distributed (i.i.d) random variables. This violation exploits the spatiotemporal dependency structure present in cities [58]. However, such dependency structures in urban data may introduce redundancy and risk an overestimation of statistical effects, it is important to take account for these redundancies, especially during the validation stages of statistical modelling.

A similarly complimentary assumption common in spatial modelling and vital within the contributions of this thesis is *stationarity*; a term used to define the (non-)uniformity of data. The stationarity assumption is used to obtain replication so that estimates can be understood by the variation of repeated observations. There are four types of stationarity; first-order, second-order, in-



(a) Second-order stationarity (b) Intrinsic stationarity.

Figure 1.2: Two example plots of second-order and intrinsically stationary processes across space: a covariogram and semivariogram respectively (see Section 2.3.1 for in-depth description).

trinsic and quasi. The most common and relevant to this thesis are: (i) second-order stationarity which implies a consistent covariance (strength of correlation) between any two pairwise observations at the same distance apart [31] (see Figure 1.2(a)) and; (ii) intrinsic stationarity which assumes that the (semi)variance of the differences between any two pairwise distances are the same (see Figure 1.2(b)). A further description of these concepts are put forth in Section 2.2.5.

In assuming stationarity, spatial models require some understanding of distance and direction. Typically, the distance function used for spatial modelling is Euclidean (also called 'as-the-crow-flies'). The Euclidean distance (more details in Section 2.4) defines a direct line between two points. This is frequently unrealistic for urban settings containing physical restrictions and social structures, for example road and path networks, large areas of private land and legal restrictions such as speed limits and one-way systems. Consequently, the features of a Euclidean distance do not take account for spatially dependant urban processes. For example, a citizen's perception of space in their own city may in fact be more related to their *perceived accessibility to place*, a concept qualitatively discussed by Neogeographers [58]. Hence, most urban processes contain some level of non-Euclidean decision making. For example, the price that an

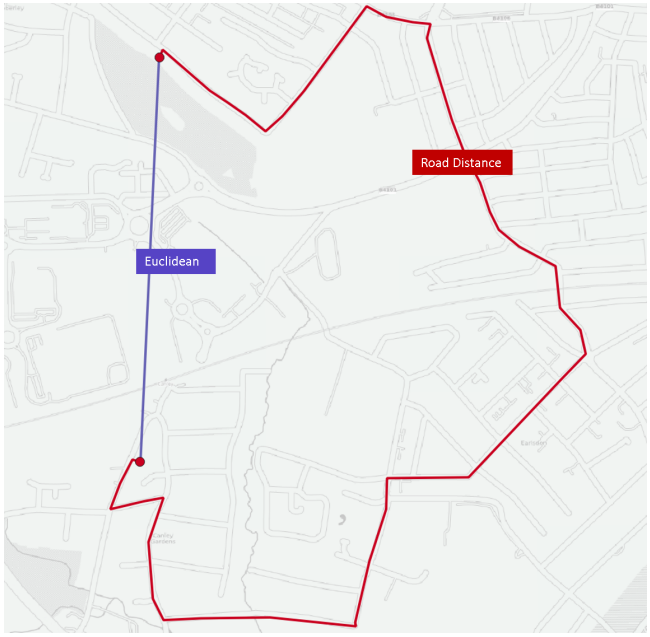


Figure 1.3: A geographical representation of a Euclidean route versus the route taken along a road network.

urban citizen is willing to pay for a house or the acceptable distance a person is happy to travel to work, parkland or shops. This thesis addresses the need for applying non-Euclidean distance approximations for geostatistical urban models (i.e., road distance and journey time). Figure 1.3 provides an example of why the proximity of two observations may differ significantly with road distance compared to Euclidean measures. At several stages throughout this thesis a set of 'motivating examples' identify the non-trivial challenges of these changes, most notably, the requirement for distance functions to lay in a metric space; a feature not apparent in road distance and travel time, see Section 2.4.

The remainder of this chapter will introduce all of the major contributions that build the scientific novelty and industrial impact of this thesis. Chapter 2 provides a full description of background research inclusive of the themes, definitions and concepts relevant to this research. Chapter 3 introduces the datasets and case studies utilised throughout this thesis. Chapter 4 presents the first major contribution which considers the use of non-Euclidean distances

for geostatistical urban models. Chapter 5 thereafter examines the second major contribution whereby the methodological implications of modelling space with non-Euclidean distances are considered. My third and final contribution is presented in Chapter 6 in which a new state-of-the-art cross validation (CV) approach for more appropriately estimating the generalisation performance of urban-specific models with spatial data is introduced. Thereafter, Chapter 7 puts forth a set of answers to the research questions (RQ) posed in Section 1.1 and discusses the implications of my work on the research area of Urban Science. Finally, Chapter 8 concludes all of my findings and puts forth a set of research avenues that are opened up by this thesis.

1.1 Research Questions and Contributions

This section introduces a set of RQs for consideration along with a brief description of the primary contributions put forth to address and answer each one individually. Each contribution is associated (and cross-referenced) to at least one publication presented in Section 1.2.

RQ1: Which distance function best models spatial interactions in an urban setting?

Urban processes in space result in data which are not i.i.d and as such semi-variograms [33], Moran's I [96] or Getis's G [56] have been put forth to statistically measure the extent of these dependencies and hence take them into account. Each of these methods have a notable commonality - distance is measured with a Euclidean function (defined in Section 2.4.1). Hence, these distance-based learning methods do not take account for physical properties of dispersion in a city landscape; for example in real estate, a person's decision to buy may consider; (1) their distance or journey time to specific locations (workplace for example) or; (2) the comparable prices of other sub-markets within close proximity. As such, I propose that physical barriers such as buildings,

roads, paths and non-accessible open space can be modelled by the distance or travel time along a (restricted) road network instead.

Contributions to RQ1: Contribution 1 (Chapter 4) considers this issue by putting forth three approximate restricted road distance, travel time and combined pairwise distance matrices to predict the value at specific house sale locations. To ensure a valid distance metric for geostatistical modelling (refer to Sections 2.3.2 and 2.4 for definitions), I propose that the Minkowski distance function with a P-value (definition in Section 2.4.1) most correlated to the OSM road network data is a strong approximation of space and proximity in the city. The work in this contribution is taken directly from Publication 4 in Section 1.2.

RQ2: When, if ever, are non-Euclidean distance functions valid for urban spatial models?

For spatial prediction (i.e., Kriging - Section 2.3.2), it is essential to ensure that existing covariance and (semi)variance functions remain valid; positive definite (PD) and conditionally negative definite (CND) respectively [39] (see full explanation in Section 2.3.1). Given the extensive work on spatial modelling with a straight line - Euclidean pairwise distance - there is no guarantee that any non-Euclidean distance matrix (PD or otherwise) will produce a valid covariance or (semi)variance function, a proof of this is provided in Chapter 5.

Contributions to RQ2: My second contribution (Chapter 5) puts forward a method to approximate restricted road distance, journey time and combined matrices using an embedded lower-dimensional Euclidean space. This method ensures that covariance and (semi)variance functions remain valid when using urban-specific distances. For confirmation, I provide a comparison of six Ordinary Kriging predictions (definition in Section 2.3.2), each with a different distance metric, employed in a real estate case study. The work in this contri-

bution is taken directly from Publication 5 in Section 1.2.

RQ3: How should one estimate the generalisation performance of urban models containing spatial data?

CV splits a dataset into two subsets: a **training set** which a model is fitted on and a **validation test set** which the model predicts and is evaluated on [117]. The main purpose of CV is to detect over fitting and estimate how well a model will **generalise to unseen data** i.e., the expected performance of a ground truth test set (defined in Chapter 6). Furthermore, k-fold cross validation (KCV) repeats the process k times while appropriately validating on all the disjoint subsets of the dataset. This method assumes that the random variables in the validation test and training set are i.i.d. However, urban problems are inherently spatial, which invalidates this assumption. As such, spatial k-fold cross validation (S-KCV) [106] attempts to remove the SAC between the training and validation test set. Specifically, S-KCV implements a Euclidean 'dead-zone' area around all test points, such that all training points that lay in these areas are removed. As per contributions 1 and 2, I propose that a non-Euclidean dead-zone will better infer the interactions contained in urban space.

Contributions to RQ3: In Chapter 6, I introduce a new spatial k-fold cross validation method, entitled road distance and travel time k-fold cross validation (RT-KCV). This method constructs *road network and travel time* dead-zones to better estimate urban SAC. I show that RT-KCV outperforms the current state-of-the-art for estimating the generalisation performance of any geostatistical urban model across the full interpolation-extrapolation range of application scenarios. The work in this contribution is taken directly from Publication 6 in Section 1.2.

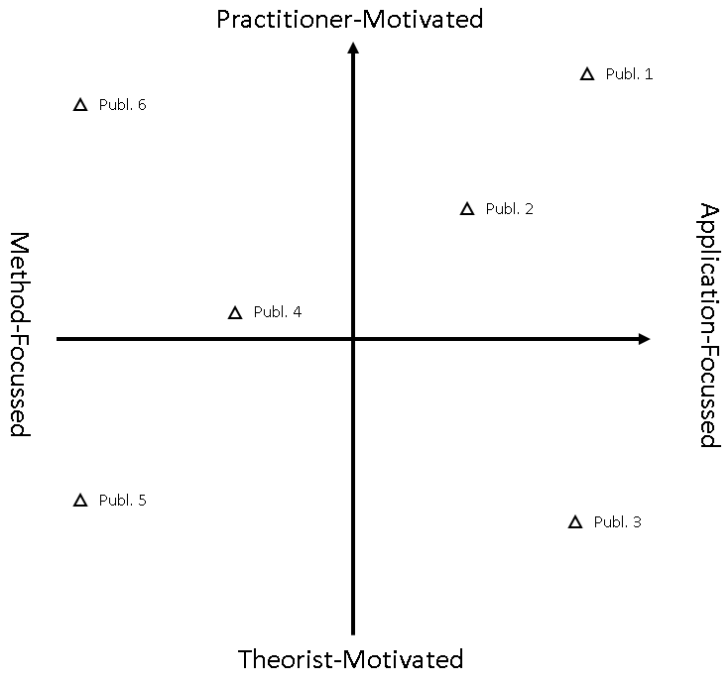


Figure 1.4: A visualisation showing the spread of research aims by contribution, related to publications. The x-axis represents motivation from practitioner to theorist and the y-axis defines focus from method to application.

1.2 Publications

Chapters 4-7 contain the work undertaken within publications 1-6. Figure 1.4 provides a visual representation of the motivations and focusses of each publication discussed. This graph shows the variety of work in this thesis.

1. Crosby, H., Davis, P. and Jarvis, S.A., 2015. Exploring New Data Sources to Improve UK Land Parcel Valuation. In Proceedings of the 1st International ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics. Published.

2. Crosby, H., Davis, P. and Jarvis, S.A., 2016, September. Spatially-Intensive Decision Tree Prediction of Traffic Flow across the entire UK Road Network. IEEE/ACM 20th International Symposium on Distributed Simulation and Real Time Applications (pp. 116-119). Published.

3. Crosby, H., Davis, P., Damoulas, T. and Jarvis, S.A., 2016, October. A spatio-temporal, Gaussian process regression, real-estate price predictor. In Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (p. 68). ACM. Published.

4. Crosby, H., Damoulas, T., Porto de Albuquerque J., Caton, A., and Jarvis, S.A., 2018, May. Road distance and travel time for an improved house price Kriging predictor, Geo-spatial Information Science (IGIS), 21:3, 185-194, DOI: 10.1080/10095020.2018.1503775.

5. Crosby H., Damoulas T and Jarvis S.A., 2019. Embedding road networks and travel time into distance metrics for urban modelling, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2018.1547386

6. Crosby, H., Damoulas, T. and Jarvis, S.A., 2018, November. Road and Travel Time Validation for Urban Modelling. International Journal of Geographic Information Sciences (IJGIS). In final draft before submission.

In addition, I have authored a number of other papers in the domain of Urban Science, which are not discussed in this thesis. These can be seen below (publications 7-10):

7. Tkachenko, N., Chotvijit, S., Gupta, N., Bradley, E., Gilks, C., Guo, W., Crosby, H., Shore, E., Thiarai, M., Procter, R. and Jarvis, S., 2017. Google trends can improve surveillance of type 2 diabetes. Scientific reports, 7(1), p.4993. Published.

8. Gupta, N., Crosby, H., Guo, W., Proctor, R., Jarvis, S., 'Twitter Usage Across Industry: a Spatiotemporal Analysis'. IEEE International Conference

on Big Data Computing Service and Applications, Germany, Mar 2018. Published.

9. Mansour, A., Crosby, H., Perera, S., Jarvis, S. Who Follows Who? A retail Agglomeration Phenomena. International Journal of Geographic Information Sciences (IJGIS). Under review.

10. Titis, E., Crosby, H., Proctor, R., Jarvis, S. Finding a golden food desert measure: examining correlations between obesity and self-derived distance measures in greater London. Under review.

CHAPTER 2

Background Research

Urban science regularly takes on new and different perspectives of the city. Each change reflectively enables a better understanding of the dynamic and multivariate nature of urban processes, as discussed in the introduction. A popular recent approach (dubbed 'The New Science of Cities' [10]) is application-motivated, however focusses on methodological contributions from multiple disciplines to process the functions, challenges and solutions of a city. Unlike previous iterations and other versions of Urban Science (Urban Social Geography [71], Urban Economics [27] and Urban Planning [48] to name a few), the latest wave of urban scientists (led by [11, 124]) focus on interdisciplinarity: combining quantitative, behavioural, structural and post-structural perspectives [71]. In the true nature of contemporary Urban Science, this thesis will promote a set of *data-driven* contributions across a combination of disciplines: GIS, Geoscience, Computer Science, Statistics, Machine Learning and Data Science. As such, Chapter 2 attempts to provide a full description of the history, themes, definitions and concepts required for each element of this interdisciplinary thesis.

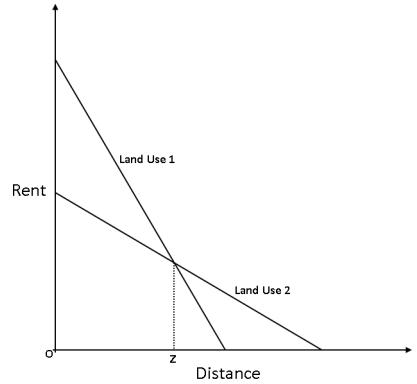
This chapter will begin with a section discussing the (spatial) complexities of the city including a history of urban space theory. Thereafter, a set of essential concepts in spatial statistics are introduced. Next, geostatistical modelling is discussed, specifically semivariograms and Kriging. Then the relevance of distance functions for spatial models and cross validation (CV) is discussed. Afterwards, there is a review of the current state-of-the-art with regards to related urban case studies; house price and traffic flow prediction. Penultimately, the practicalities of storing, retrieving and analysing spatial data is explored in detail, before offering some final remarks.

2.1 A History of Urban Space Theory

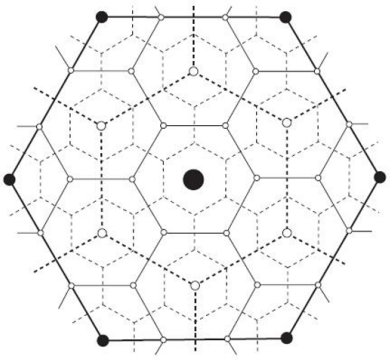
Pre-dating central Europe's industrial revolution, and hence major urbanisation [59], is the domain of urban space theory, conceived by *land economists* who attempted to understand the spatial interactions motivating the design of a city. Consistently, each theory in this domain describes (1) *land rent* as being the key driver for design and (2) *space* as being the key influencer to rent, which is a concept grounded by that of Von Thünen's *Isolierte Staat* ("Isolated State")[123].

Von Thünen's study examined the patterns of agricultural land surrounding the 19th century city. He put forth that the primary function pertaining to agricultural competition was *economic rent* (i.e., the land's return on investment (ROI)), in which transportation costs were the primary factor. This function is visualised in Figure 2.1(a), where two land uses are considered: *land use 1* is less desirable than *land use 2* at all distances greater than Z. These differing gradients are due to the transportation costs for the specific produce to be sourced on that land and taken to the city i.e., the produce weight, refrigeration requirement or the vehicle type needed to transport the produce. Von Thünen's work contained a number of assumptions inappropriate for the contemporary city; (1) a static (non-changeable) space, which (2) fully supplies a single market centre with (3) no competition. Indeed, urban sprawl and economic competition in the 21st century are not so simplistic. They can be based on the availability of land, the adoption of local market centres and complex topographies such as rivers, extreme elevation and valleys.

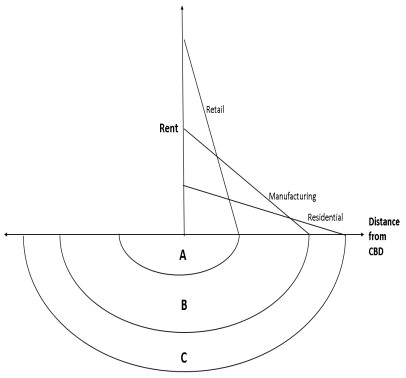
In an attempt to address some of these issues, the Central Place Theory (CPT) instead describes an urban community as being a system of multiple central places (towns and cities) whose primary functions are to be mediators for local commerce [21]. CPT assumes an urban system to be a single large community spatially contained by a number of smaller ones (towns, villages and hamlets). The theory puts forth a hierarchy of market centres in which



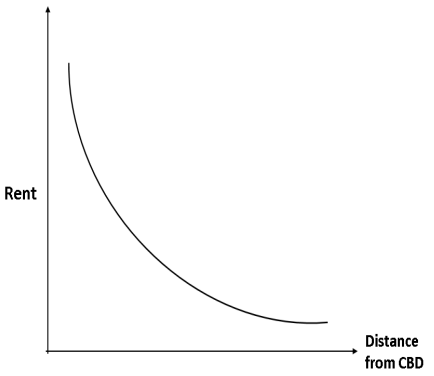
(a) Relationship of economic rent and distance from the market centre for two competing land uses [123].



(b) A visual description of urban space using the Central Place Theory with four orders of magnitude [21].



(c) A visual description of the Bid-Rent theory for three land use types [1].



(d) The Bid-Rent theory showing non-linear diminishing returns between the distance from a CBD and rent [1].

Figure 2.1: A comparison of three urban space theories: The Isolated State [123], Central Place Theory [21] and Bid-Rent Analysis [1].

the larger ones offer goods and services that are not supported by their smaller counterparts. In his own study, he notes 7 layers of hierarchy where all centres are equidistant from one and another; this does however assume that settlement patterns are uniform. Figure 2.1(b) shows an example with five levels of hierarchy. CPT assumes no boundary, a homogeneous plain of land uses and no topographical constraints. In his later work, Christaller would more appropriately go on to note the importance of population distributions to urban centres [21]. Both of the aforementioned models inspect but do not explain the functions and flows of urban centres [112].

Another key contribution in the area of traditional urban land theory is Alonso's bid-rent model which more appropriately proposes that the value of land (and hence the design of the community that it is supporting) is multivariate i.e., a function of transportation costs, land use, land intensity, population and employment [1]. His study introduced two concepts which are now largely relevant to contemporary urban modelling and my thesis in particular: (1) the land owner's settlement location is multivariate; and (2) a city's design depends on each sector of urban real estate (retail, manufacturing and residential). Each sector have significantly different *utility* functions i.e., each variable determining a lands rent differs between each individual stakeholder. Figure 2.1(c) visualises this theory whereby a land owner's elasticity to space (responsiveness to demand relative to the land's distance from the market centre) is much stricter in the retail sector than in manufacturing and residential [87]. Hence, land use within cities split into *zones* such that zone A is primarily retail, zone B is manufacturing and zone C is typically residential. Furthermore, Figure 2.1(d) shows that the granular relationship between distance and rent, in Alonso's model, is actually non-linear. This is due to the fact that some features may be more influential to a land purchaser's utility than transportation costs, for example increased land size (more apparent in the suburbs) may be disproportionately more desirable to a residential land owner than their distance to a market centre.

The use of space in each of these theories ground and motivate the work put forth within this thesis. Most notably, 'proximity' as a driver for urban systems proves particularly relevant. Contemporary discussions, on the other hand, are facilitated by a number of formalised concepts in the area of spatial analysis, specifically stationarity (Section 2.2.5), stochastic processes (Section 2.2.2) and spatial autocorrelation (SAC) (Section 2.2.4).

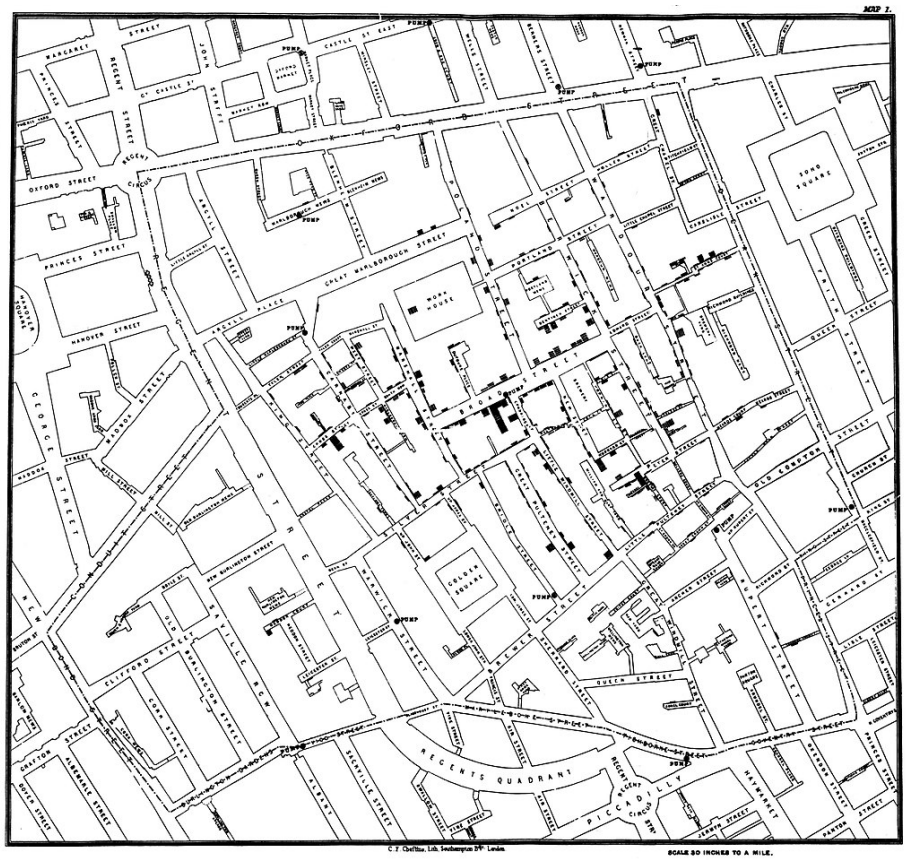


Figure 2.2: The London cholera map produced by Dr John Snow. Accessed from *http://www.ph.ucla.edu/epi/snow/snowmap1.pdf* on 06 June 2018.

2.2 Spatial Analysis

Spatial Analysis is the research area which provides a set of methods for analysing interactions in geographical space. This entire paradigm is centred around the concept of proximity.

2.2.1 A motivating example

Dr John Snow's iconic cholera map (see Figure 2.2) motivates some of today's most popular urban space analysis theories and models. His map identified the causes and geographical origins of a cholera epidemic in the Soho district of London. The 1854 map reported cholera cases as black rectangles, where

more rectangles represent a higher number of cases. The map shows variation which identifies spatially dependant variables (stochastic factors - see definition in Section 2.2.2), such as population density and water supplier. The map also shows that the number of observations in an area interact with their immediate neighbours, a concept now grounded in contemporary proximity analysis i.e., stationarity (Section 2.2.5) and SAC (Section 2.2.4).

In addition, his map recognised the effect of the built environment on distance (i.e., the distance of cholera cases along a road network), which is a concept central to this thesis but rare in contemporary spatial analysis. Incidentally, he concluded that the source of cholera was a water pump on Broad Street, unlike popular believe, which assumed cholera to spread through the air. The remainder of this section will introduce the aforementioned key concepts in spatial analysis.

2.2.2 Spatial stochastic processes

Cressie's book entitled *Statistics for Spatial Data* comments that *"statistics ... attempts to model order in disorder"*, an apt starting premise for this section and the remainder of this thesis [31]. In fact, with the right statistical models, the behaviours of disordered or *random* variables, despite their name, can be structured and predictable i.e., random variables are described by *probability spaces*.

Definition 2.2.1. *Probability spaces and random variables.* A probability space is a random process or experiment with components (Ω, F, P) where Ω is a sample space of possible outcomes (O), F is a set of possible events (E) and P is a probability measure over Ω. Any mapping $Z : \Omega \mapsto F$ is called a random variable, whereby F is a measurable space of E with respect to P. The probability measure $P : \omega \mapsto [0,1]$ assigns probabilities (used as weights) to individual outcomes $\omega \in \Omega$ and also allows the assessment of the probability of events $E \in F$. When random variables are referenced over an additional structure, known as

an index set, a sequence of random outcomes called *stochastic processes* are witnessed [28].

Definition 2.2.2. *Stochastic processes.* Data point s varies over index set $D \subset \mathbb{R}^d$ if (1) $s \in \mathbb{R}^d$ is a generic data location in d-dimensional Euclidean space, (2) $Z_s \equiv Z(s)$ is a potential datum at spatial location s and (3) s is a random value. As such, a collection of random variables generating a (multivariate) random field is called a stochastic process, such that

$$Z = \{Z(s) : s \in D, Z \in \Omega\}. \tag{2.1}$$

Typically, D is assumed to be a fixed (non-random) subset of $\mathbb{R}^d$ where stochastic processes represent time-series data with T (instead of D) and t (instead of s) often denoting a time interval $[t_0, t_n] \subset \mathbb{R}$. However, given that D is a subset of $\mathbb{R}^d$ then Equation 2.1 is a spatial stochastic process. For completeness, a time-series process is

$$\{Z(t) : |t| < \infty\} \tag{2.2}$$

and a space-time process is

$$\{Z(s;t) : s \in D, t \in T\} \tag{2.3}$$

where Z, D and T are all random. If one now assumes that $T = \mathbb{R}^2$ is a two-dimensional real-valued index set and $S \subset T$ is a set of spatial units then three different kinds of spatial processes are obtained [31]: spatial random fields, lattices and spatial point patterns [31]. Each of these processes produces a 'type' of spatial data; geostatistical data, lattice data and point process data respectively.

2.2.3 Types of spatial data

Section 2.2.2 determined that spatial data can result from observations on the stochastic process $Z = \{Z(s) : \mathbf{s} \in D, Z \in \Omega\}$, where D is a random set in $\mathbb{R}^d$.

This spatial data is likely to be one of three types; *geostatistical, lattice* or *point patterns.*

Definition 2.2.3. *Geostatistical data.* A collection of random variables $Z = \{Z(s) : s \in D, Z \in \Omega\}$ is called a spatial random field if the spatial index set is continuous and fixed. For example, D is a fixed subset of $\mathbb{R}^d$, $Z(s)$ is a random vector at location $s \in D$ and $|s| = \infty$.

A more descriptive interpretation is that geostatistical data are selected points within a spatial process containing continuous variation. A commonly referenced example is a dataset of mineral concentrations (a spatially continuous variant) which are sourced at specific drilling locations. This data is usually analysed within the domain of Geostatistics [88]; a science which recognises variation on both large and small scale areas and observes both spatial trends and spatial correlations.

Definition 2.2.4. *Lattice Data.* A collection of random variables $Z = \{Z(s) : s \in D, Z \in \Omega\}$ are called lattice data if the spatial index set is discrete and fixed. For example, D is a fixed (regular or irregular) collection of frequently countable points in $\mathbb{R}^d$ (i.e., $|s| < \infty$), D is a graph in $\mathbb{R}^d$ and $Z(s)$ is a random vector of all locations $s \in D$.

Intuitively, lattice data are locations that observe spatial processes at regular (or irregular) grids. This data is usually sourced from some spatially aggregated area. Commonly referenced sources of lattice data are satellite surveys reporting average weather patterns, land heights or crop distributions across small aggregated areas. Typically, this data is utilised in the research area of Remote Sensing where analysis is usually large scale.

Definition 2.2.5. *Point Patterns.* Spatial units $S = \{S_i \in \mathbb{R}^2 : i \in \mathbb{N}\}$ are point processes if they are random variables. For example, D is a point process in $\mathbb{R}^d$ or a subset of $\mathbb{R}^d$ and $Z(s)$ is a random vector at location $s \in D$.

By way of explanation, spatial point processes are regularly or irregularly

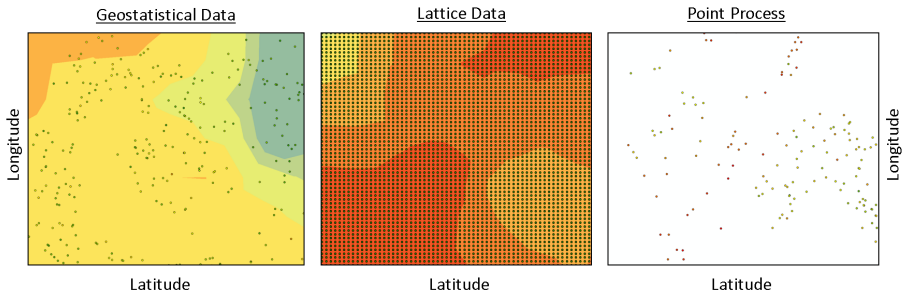


Figure 2.3: A visual representation of three types of spatial data: geostatistical, lattice and point process.

spaced locations of interest which are measured at their exact location. Popularly discussed examples of point patterns are galaxies in space or trees in a forest, which may contain no continuous structure, but instead contain other relationships such as clusters. Figure 2.3 provides a simple visualisation for each of the discussed data types.

There is a fourth and irregularly referenced type of spatial data named an 'object' where D is a point process in $\mathbb{R}^d$ and Z(s) is itself a random set. This data type is not discussed any further in this thesis due to it's lack of relevance to my work.

The importance of understanding your data

It is essential to understand an application's 'data type' before making any conclusions due to the ambiguity of some processes. For example, a dataset could appear to be a spatial point process however the observations may actually interact along continuous space (i.e., a random field). My primary case study (house prices) reflects this exact problem, as the random fields present in house price prediction (shown in Chapter 4) gives rise to geostatistical analysis, despite first appearance, where one might assume point process data instead. Hence, the work from this thesis will focus on utilising, challenging and improving the methods put forth in the domain of Geostatistics, as will the following sections

of this Chapter.

2.2.4 Spatial autocorrelation (SAC)

The first law of geography states that "everything is related to everything else, but near things are more related than distant things" [92]. This violates the assumption used by non-spatial statistics: that observations are independent and identically distributed (i.i.d) random variables [43]. Observations as i.i.d random variables is a reasonable assumption when samples are taken from controlled experiments containing no interactions, however dependency structures in space explain the effect that geographical proximity has on data points in a spatial distribution [66, 92]. This concept is defined as SAC. An example of such SAC could be rainfall or the spread of airborne diseases.

Formally, SAC shows that observed attributes of closer points are more similar than those that are further from each other [45, 79]. When modelling random variables, SAC may need to be taken into account for a multitude of reasons, notably to: test on model misspecifications [24]; measure the strength of spatial effects on a variable; test for spatial stationarity, heterogeneity or clustering (see section 2.2.5 for more details on these concepts); detect distance decay; identify outliers and design spatial samples [2, 50, 55].

There are multiple ways to measure SAC; Moran's I, Getis-ord's G_i^* and Matheron's (semi)variogram are the most popular. The Moran's I-test measures the relationship between the lag of pairwise points and the covariance of observations; this statistic assumes and measures second order stationarity (see Section 2.2.5). Getis-ord's G_i^* statistic [56] reports the location of observations that are clustered spatially for extreme values (described as hot or cold-spots). Finally, Matheron's (semi)variogram measures the (semi)variance of the differences between any two pairwise distances; this statistic assumes and measures intrinsic stationarity and will be discussed in Section 2.2.5).

Definition 2.2.6. *Moran's I-Test.* Moran's I [96] estimates the normalised spatially-weighted covariances of all random variables. The spatial weights de-

fine the fixed geographic structure connecting spatial units $s_i \in S$. The covariance matrix is commonly calculated giving a weight of 1 to its k nearest neighbours and 0 otherwise. Moran's I takes account of pairwise relationships between spatial units by using a spatial weights matrix W). The global form of Moran's I, which we consider in this thesis, averages the overall SAC in a region and is defined as:

$$I = \frac{N}{s_0} \frac{\sum_i \sum_j w_{ij}(s_i - \bar{s})(s_j - \bar{s})}{\sum_i (s_i - \bar{s})^2} \tag{2.4}$$

where s_i are the observations, $w_{i,j}$ are the distance weightings, N is the number of observations and $s_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}$. If $I_{observed} >> I_{expected}$ then the values of s are positively autocorrelated else they are weakly or negatively correlated [96].

Definition 2.2.7. *Getis-Ord's G_i^*-Statistic.* G_i^* is an extension to the popularly employed general G-statistic [57]. The test measures for spatial patterns. Unlike the global Moran's I-test discussed above, the G_i^* statistic assumes that spatial dependency may vary significantly over a study area (i.e., assumes local non-stationarity - defined in Section 2.2.5). Specifically, the G_i^*-statistic measures localised extremal values (i.e., hotspot and coldspot clusters).

Formally, the G_i^* statistic is

$$G_i^* = \frac{\sum_{j=1}^{n} w_{ij} \cdot s_j}{\sum_{j=1}^{n} s_j}, \quad \text{for } j \neq i \tag{2.5}$$

where w_{ij} is a weight value between event i and j that represents their spatial interrelationship and s_j is the magnitude of variable S at incident location j over all n where $j \neq i$. Usually, w_{ij} is calculated based on the maximum conceptualized distance of spatial relationships (i.e., a user-specified distance threshold). G_i^*-statistic is commonly used to locate the best place for emergency services or taxi pick-ups in a city. The primary difference between Moran's I and the G_i^* statistics is the way that the weightings are calculated. Most

importantly, the global Moran's I-test tells you how much clustering and SAC there is whereas the Getis-Ord's G_i^*-statistic shows you where the clusters are.

2.2.5 Spatial stationarity

The stationarity assumption is used to obtain replication so that estimates can be understood by the variation of repeated observations. Formally, $Z(s)$ is stationary if, for any finite number of n points $s_1, \ldots, s_n$ and any distance h, the joint distribution of $Z(s_1), \ldots, Z(s_n)$ is the same as the joint distribution of $Z(s_1 + h), \ldots, Z(s_n + h)$ [28]. The three most commonly considered types of stationarity, each with different degrees of constraint are first-order, second-order and intrinsic.

First-order stationarity assumes that the data's mean average (first order moment) is constant over space and *second-order* stationarity assumes that the mean and covariance (first and second-order moments) are constant over space. In first and second-order stationarity, the covariance is dependent only on distance and not location i.e., it is a function of lag only. All higher ordered moments i.e., covariance and kurtosis for first order and kurtosis for second order contain variation. *Intrinsic* stationarity assumes that the expected values of the mean and variance are constant with respect to location (i.e., in all directions). The models considered in this thesis either assume second-order [26] or intrinsic [57] stationarity.

Definition 2.2.8. *Second-Order Stationarity.* Given a finite set (D) of n discrete spatial points s, one can obtain a stochastic process $Z = \{Z(s) : s \in D, Z \in \Omega\}$ with $s, s_1, s_2 \in S$ and $E[Z(s)^2] < \infty$ for all s. This process is second-order stationary iff

$$E[Z(s)] = \mu, \tag{2.6}$$

$$Var[Z(s)] = \sigma^2 \tag{2.7}$$

and

$$cov[Z(s_1), Z(s_2)] = \sigma^2 \cdot C(||s_1 - s_2||) \text{ for all } s_1, s_2 \in D. \tag{2.8}$$

$C(||s_1 - s_2||)$ is the correlation function and $||s_1 - s_2||$ is the Euclidean norm. The resulting process contains a covariance reducing to a function of distances. This process hence provides a consistent SAC for statistical tests.

Definition 2.2.9. *Intrinsic Stationarity.* Let D be a finite set of n discrete spatial points s. Further, consider set $Z = \{Z(s) : s \in D, Z \in \Omega\}$ of spatial random variables. Then *intrinsic stationarity* is defined through first order differences [31]:

$$E[Z(s + \mathbf{h}) - Z(s)] = 0 \tag{2.9}$$

and

$$Var[Z(s + \mathbf{h} - Z(s))] = 2\gamma(h) \tag{2.10}$$

where $\mathbf{h}$ is a specific lag and $2\gamma(\mathbf{h})$ is a variogram [88] which is defined in Section 2.3.1.

2.3 Modelling Random Fields

As previously discussed in Section 2.2.3, Geostatistics is concerned with modelling geostatistical data from random fields. These are discrete data points that represent a single location on a continuous plane across a spatial region. This is achieved through geostatistical modelling which always rely on semivariograms to obtain the spatial dependency of target data.

2.3.1 Semivariogram (γ) / variogram (2γ)

A (semi)variogram describes the spatial relationships between all observations measured at a (typically omnidirectional) distance and assumes the input dataset to be intrinsically stationary, as defined in Section 2.2.5.

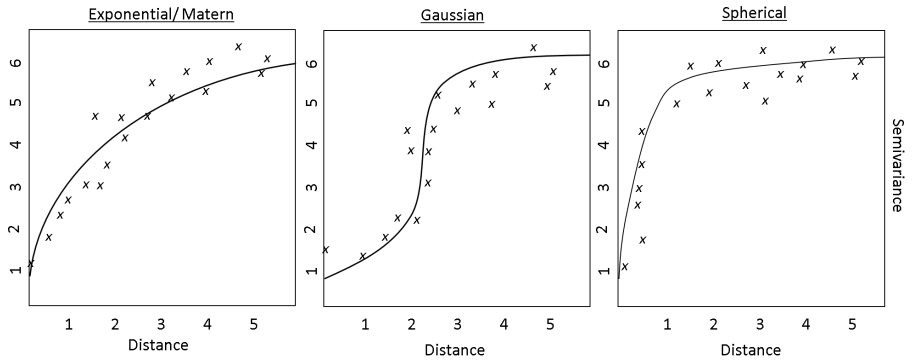


Figure 2.4: Semivariogram kernel examples; exponential/Matern, Gaussian and spherical.

First, the empirical (semi)variogram is calculated by finding the average (semi)variance at a set of user-defined lags - the points in Figure 2.4 represents an empirical semivariogram. Thereafter, a model named the experimental (semi)variogram (also known as the kernel or covariance matrix) is selected (empirically or otherwise) based on a (typically parametric) function that best fits the empirical (semi)variogram, such as; Gaussian, Matern, spherical, exponential and so on. Each of the aforementioned parametric functions have been designed to best infer (semi)variance based on Euclidean geographical proximity. Figure 2.4 provides an example of three commonly applied experimental semivariograms; exponential, Gaussian and spherical.

A set of hyperparameters are selected to calculate the experimental semivariogram; nugget, sill and range. The nugget is the value at which the (semi)variogram is very close to 0 (i.e., almost intercepts the y-axis). The sill and range are the variance and distance (respectively) at which the gradient of the variogram (γ) becomes 0. Each of these concepts are visualised in Figure 2.5. It can be seen from this figure that (semi)variograms can provide some measure of distance decay, which has potential to inform non-global spatial models. Importantly, a small semivariance implies a strong pairwise relationship between observations.

Formally, let the variance between two observed locations s_i and s_j be: $var(Z(s_i) - Z(s_j)) = 2\gamma(s_i - s_j)$, for all $s_i, s_j \in D$. The variogram is $2\gamma(\mathbf{h})$

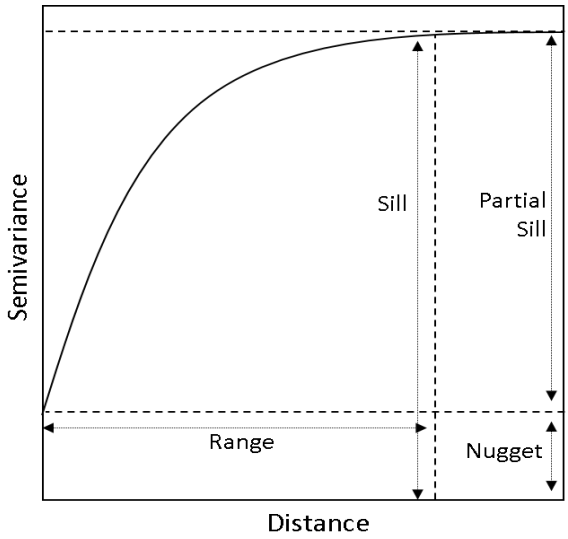


Figure 2.5: A visual description of variogram parameters; nugget, range and sill.

(a function of the increment $s_i - s_j$) and the semivariogram is $\gamma(h)$ [88]. The function $2\gamma(\mathbf{h})$ (assuming it exists - see Section 5.3) is a parameter of the random process $Z(\cdot)$ defined as

$$2\gamma(\mathbf{h}) = \frac{1}{N(\mathbf{h})} \cdot \sum_{N(\mathbf{h})} (Z(s_i) - Z(s_j))^2, \tag{2.11}$$

where $N(\mathbf{h}) \equiv \{(i,j) : s_i - s_j = \mathbf{h}\}$ and $|N(\mathbf{h})|$ is the number of distinct elements of $N(\mathbf{h})$.

As with almost any parametric function, variograms contain some constraints, notably the requirement to be conditionally negative definite (CND), i.e.:

$$\sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j 2\gamma(s_i - s_j) \leq 0 \tag{2.12}$$

for any finite number of spatial location $\{s_i : i = 1, \ldots, m\}$. Where the real numbers $\{a_i : i = 1, \ldots, m\}$ satisfy $\sum + i = 1^m a_i = 0$. As proof, lets suppose for the moment that $Z(\cdot)$ is an intrinsically stationary process (i.e., it has constant

mean and possesses a variogram $2\gamma(h)$), then

$$\{\sum_{i=1}^{m} a_i z(s_i)\}^2 = -(\frac{1}{2}) * \{\sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j (Z(s_i) - Z(s_j))\}^2 \tag{2.13}$$

because $\sum_{i=1}^{m} a_i = 0$. Upon taking expectations, one obtains:

$$\sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j 2\gamma(s_i - s_j)\} = 2 \text{ var}\{\sum_{i=1}^{m} a_i z(s_i)\} \leq 0 \tag{2.14}$$

hence the function $2\gamma(h)$ is CND when the stochastic process is intrinsically stationary (extensive proof by Cressie, pg. 87 [33]). Given that the data analysed in this thesis are spatial random fields which are typically assessed by geostatistical research, a variogram will be essential for modelling, the reasons for this become apparent in the next chapter.

2.3.2 Kriging

Kriging is a geostatistical spatial predictor which accounts for spatial covariance based on observation distances to understand the spatial structure of a dataset and hence determine it's regression parameters. Kriging is used extensively for interpolation by Ecologists [81], Geographers [19] and Geoscientists [67]. The basis of Kriging is to first model the degree to which distance between observations is correlated using the experimental variogram and then apply modelling coefficients to determine interpolation parameters based on the spatial patterns determined by the variogram.

Formally, Kriging serves to estimate the value $Z(s_0)$ at point s_0 with a known variogram conducted by the neighbouring points of s_0. The way in which the interpolation weights are calculated determines the 'type' of Kriging undertaken; Simple, Ordinary or Universal to name the most popular. The selected method should be based on the stochastic properties of the random field studied and the type of stationarity assumed.

Simple Kriging assumes first-order stationarity across the whole region with

a known mean m, i.e., $E\{Z(s)\} = E\{Z(s_0)\} = m$. *Ordinary Kriging* assumes a constant unknown mean over the variogram-defined neighbourhood for s_0. Finally, *Universal Kriging* assumes a general trend of any polynomial order. Throughout this thesis, I exclusively utilise Ordinary Kriging, the reasons for this are explained at the times of implementation in Chapters 4-6.

Definition 2.3.1. *Ordinary Kriging.* Ordinary Kriging implicitly evaluates the mean of a moving neighbourhood. This is only valid when (1) the dataset is intrinsically and second-order stationary and (2) an experimental semivariogram $\gamma(\mathbf{h})$ is calculated and present [125]. Generally, a Kriging estimator of the local mean is set up, then a simple estimator is taken from the Kriged mean. To estimate $Z(s_0)$ at location s_0, the data values $Z(s_i)$ from n neighbouring sample points are multiplied by some linear weights λ_i, such that:

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i). \tag{2.15}$$

Notably, $\sum \lambda_i = 1$ so that, in the case where all of the $Z(s_i)$ values are a single constant, the estimated value $Z(s_0)$ must be equal to that same constant. This guarantees uniform unbiasedness (Equation 2.9). Importantly, the model assumes the data to be part of a realisation of an intrinsic random function with $\gamma(\mathbf{h})$. Given that the expectation of each increment is 0, unbiasedness with unit sum weights is calculated:

$$\begin{aligned} E[\hat{Z}(s_0) - Z(s_0)] &= E[\sum_{i=1}^{n} \lambda_i Z(s_i) - Z(s_0) * \sum_{i=1}^{n} \lambda_i] \\ &= \sum_{i=1}^{n} \lambda_i E[z(s_i) - z(s_0)] = 0. \end{aligned} \tag{2.16}$$

The optimal Kriging predictor is then calculated by minimising the mean-squared prediction error ($\sigma^2(s_0) = E[(\hat{Z}(s_0) - Z(s_0))^2]$) over the class of linear predictors $\sum_{i=1}^{n} \lambda_i = 1$, such that $2\gamma(\mathbf{h}) = var(Z(s+h) - z(s)), h \in \mathbb{R}^d$. By minimising (i.e., differentiating and equating to 0) Equation 2.17 with respect to $\lambda_1, \ldots, \lambda_n$ and the Lagrange multiplier m (ensuring $\sum_{i=1}^{n} \lambda_i = 1$) we can

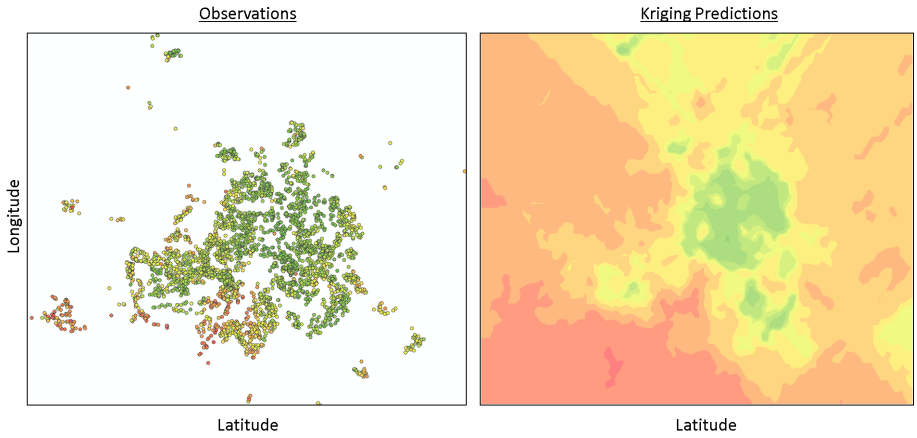


Figure 2.6: A visual representation of an input geostatistical point set (left) and a Kriged output (right).

obtain the optimal $\lambda_1, \ldots, \lambda_n$ from $\lambda_0 = \Gamma_0^{-1}\gamma_0$.

$$E\big(Z(s_o) - \sum_{i=1}^{n} \lambda_i Z(s_i)\big)^2 - 2m\big(\sum_{i=1}^{n} \lambda_i - 1\big) \tag{2.17}$$

These optimal values hence allow the provision of an Ordinary Kriging system:

$$\begin{pmatrix} 0 & \gamma(h_{12}) & \gamma(h_{13}) & \ldots & \gamma(h_{1n}) & 1 \\ \gamma(h_{12}) & 0 & \gamma(h_{12}) & \ldots & \gamma(h_{2n}) & 1 \\ \gamma(h_{32}) & \gamma(h_{32}) & 0 & \ldots & \gamma(h_{3n}) & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots & 1 \\ \gamma(h_{n3}) & \gamma(h_{n2}) & \gamma(h_{n3}) & \ldots & 0 & 1 \\ 1 & 1 & 1 & \ldots & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \ldots \\ \lambda_n \\ m \end{pmatrix} = \begin{pmatrix} \gamma(h_{01}) \\ \gamma(h_{02}) \\ \gamma(h_{03}) \\ \ldots \\ \gamma(h_{0n}) \\ 1 \end{pmatrix}$$

The weights λ_i are assigned to $Z(s_i)$, this calculation shows the disparity between all data points $Z(s_i, s_j)$:$1, \ldots, n$ (LHS) and each data point $Z(s_i)$ compared with $Z(s_0)$ (RHS). Figure 2.6 provides an example of an input dataset (LHS) and the output Kriging prediction for those observations (RHS). The red points represent high values and the green points display low values in my simulated dataset. The same colours apply for the Kriged output on the RHS.

2.4 Distance and Proximity

It has now been shown on several occasions throughout Chapters 1 and 2 that attribute values measured on features near to each other are typically more similar than those measured further apart. Given this, spatial analysis requires some measure of distance (or closeness) of which there are several ways to do so. These will be discussed in the following section.

2.4.1 Distance functions

Definition 2.4.1. *Distance Function.* D is a set of observations with spatial locations $s \in D$, also $d(s_1, s_2)$ is a real valued function representing the distance function on $D \times D$ such that $d : D \times D \to [0, \infty)$.

The distance function (also known as a distance metric) $d(\cdot) \equiv d(s_1, s_2) \equiv d_{1,2}$ measures the closeness of two arbitrary points. The most common distance functions considered in the area of geostatistics are; Euclidean [36], Manhatten [39], great arc [6] and Minkowski [35].

The Euclidean distance measures the straight-line distance between two points. The Manhatten distance measures the sum of the absolute differences between two points. The great arc distance attempts to estimate the Earth's surface by measuring the distance along a sphere. Finally, the Minkowski distance is a generalisation of both the Euclidean and Manhatten distances in a normed vector space. Figure 2.7 visualises the great arc distance and the unit circles for Manhatten, Euclidean and Minkowski (P=0.5,1.5,4,∞) distances for comparison. Of these distance functions, this thesis exclusively examines Minkowski distances, specifically the special cases of Euclidean and Manhatten (p=2, 1 respectively).

Definition 2.4.2. *Euclidean Distance.* Unless stated otherwise, it is typical to assume a Euclidean function when referring to distance, this assumption will remain valid throughout my thesis.

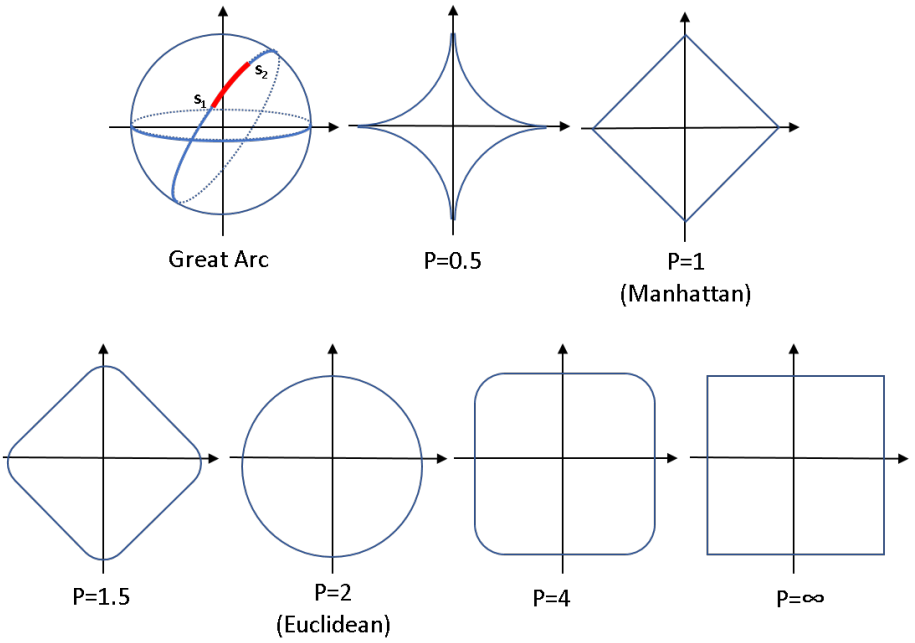


Figure 2.7: A visual representation of popularly employed distance functions.

Formally, let's assume two sites as vectors $\mathbf{s}=(s_1,\ldots,s_d)^{\mathbf{T}}$ and $\mathbf{u}=(u_1,\ldots,u_d)^{\mathbf{T}}$ in Euclidean space $\mathbb{R}^d$, hence the Euclidean distance is

$$||\mathbf{s}-\mathbf{u}|| = \{\sum_{i=1}^{d}(s_i-u_i)^2\}^{\frac{1}{2}} \tag{2.18}$$

where d is the number of dimensions (or attributes) and s_i, u_i are the attributes.

Definition 2.4.3. *Manhatten Distance.* Given the same notation as above, the Manhatten distance is

$$||\mathbf{s}-\mathbf{u}|| = |\sum_{i=1}^{d}(s_i-u_i)|. \tag{2.19}$$

The Manhatten distance is always greater than or equal to a Euclidean distance. If the locations are in a single dimension then Manhatten is always equal to Euclidean.

Definition 2.4.4. *Minkowksi Distance.* Also assuming the same notations as above, the Minkowski distance is

Figure 2.8: A two-dimensional representation of Euclidean (purple), Minkowski (green) and Manhatten (red) distances between two points.

$$||\mathbf{s} - \mathbf{u}|| = \{\sum_{i=1}^{d} ||(s_i - u_i)^P\}^{\frac{1}{P}}. \tag{2.20}$$

where P is a user defined parameter. Euclidean and Manhatten are special cases of Minkowski with values of P=2 and P=1 respectively. Figure 2.8 provides an intuitive two-dimensional example of these distance functions, where the blue line represents a Euclidean distance, the red shows a Manhatten distance and the green line shows a Minkowski distance with $1 < P < 2$ between the two red points.

In addition to these previously discussed distance functions, this thesis measures less popularly employed distances; road distance, travel time and a combination of both. Each of these distances will be introduced with the data in Chapter 3. These distances, however, are not 'functions'/'metrics'.

2.4.2 Distance metrics

For the most part, spatial statistics relies on a certain assumption; each set of distances lay in a metric space $(\mathbf{M}, d)$ where $\mathbf{M}$ is a set and d is a metric (or 'function') on $\mathbf{M}$. For example, d is a function

$$d : MXM \rightarrow \mathbb{R}^{+} \tag{2.21}$$

where $\mathbb{R}^{+} \in \mathbb{M}$ is a set of non-negative real numbers whose values satisfy requirements P1-P4:

$$d_{i,j} > 0 \text{ (P1: non-negativity)}$$
$$d_{i,j} = 0 \iff x_i = x_j \text{ (P2: identity of indiscernibles)}$$
$$d_{i,j} = d_{j,i} \text{ (P3: symmetry)}$$
$$d_{i,j} < d_{i,k} + d_{k,j} \text{ (P4: triangle inequality).}$$

A metric space is ordered so that a subset distance can be accurately measured. Although each property (P1-P4) is necessary, they are not exclusively sufficient. The 'non-negativity' and 'identity of indiscernibles' constraints define a positive definite (PD) function. Non-negativity alone defines a positive semi-definite (PSD) function. A distance satisfying PD and symmetry is called a 'semimetric' and a distance function that is PD only is called a 'divergence'. An $N \times N$ pairwise table is called a distance matrix if the metric conditions are not satisfied and a distance metric if they are.

2.4.3 Metric or matrix

It is common knowledge that the Euclidean distance function is indeed a metric. This section will provide the proof of this fact by means of example.

P1: Non-negativity

$i \neq j \implies \exists k \in [1, \ldots, n] : x_k \neq y_k$. Assuming that
$d_k(x_k, y_k) > 0$, then

$$\sum(\{d_i(s_i,u_i)^2\}^{\frac{1}{2}} > 0 \implies d(s,u) > 0.$$

P2: Identitiy of indiscernability

$$d(s,s) = \{\sum_{i=1}^{d}(s_i - s_i)^2\}^{\frac{1}{2}} = \{\sum_{i=1}^{d} 0^2\}^{\frac{1}{2}} = 0.$$

P3: Symmetry

$$d(s,u) = \{\sum_{i=1}^{d} d_i(s_i,u_i)^2\}^{\frac{1}{2}} = \{\sum_{i=1}^{d} d_i(u_i,s_i)^2\}^{\frac{1}{2}} = d(u,s).$$

P4: Triangle inequality

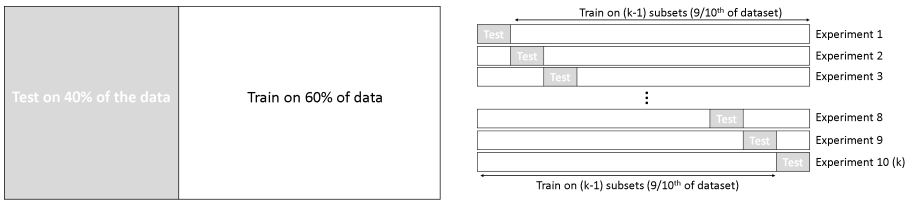
$$\text{Let } k = (k_1, k_2, \ldots, k_n) \text{ and } d_i(s_i,u_i) = \rho_i \text{ and } d_i(u_i,k_i) = \tau_i.$$
$$\text{To show that: } \sum(\{d_i(s_i,u_i)^2\}^{\frac{1}{2}} + \sum(\{d_i(s_i,k_i)^2\}^{\frac{1}{2}} \geq \sum(\{d_i(s_i,k_i)^2\}^{\frac{1}{2}},$$
$$\text{we have: d(i,j)+d(j,k)= } \{\sum \rho^2\}^{\frac{1}{2}} + \{\sum \tau^2\}^{\frac{1}{2}}$$
$$\geq \{\sum(\rho_i + \tau_i)^2\})^{\frac{1}{2}} \geq \{\sum(d_i(s_i,k_i)^2\}^{\frac{1}{2}}) = d(s,k).$$

The same intuition applies for Minkowski in all cases where $P \geq 1$, however it can be quickly shown that any $P < 1$ violates P4 (the triangle inequality). For example, if $i = (0,0)$, $j = (1,1)$ and $k = (0,1)$ in $\mathbb{R}^2$. Then for $p < 1$, $d(i,j) = 2^{\frac{1}{P}} > 2$ and d(i,k)=d(j,k)=1. Hence $d(i,j) > 2 > d(i,k) + d(j,k)$ and therefore P4 is violated.

This same approach can be taken for any distance matrix to confirm whether or not it is a metric. The purpose for determining whether a distance is a metric is very important in this thesis and discussed in great detail in Chapters 4 and 5.

2.5 Validating models with spatial data

The primary aim of model validation is to report the performance of a model. There are two stages to model validation: selecting a *cross validation* method and selecting a *validation metric*.



(a) Holdout. (b) k-fold.

Figure 2.9: Holdout versus KCV techniques.

Cross validation

Cross validation (CV) splits the dataset into two subsets: a training set where a model is fitted on and a validation test set where the model predicts and is evaluated on [117]. The main purpose of CV is to detect over-fitting and estimate how well a model will generalise to unseen data. The most common and reliable CV techniques are holdout and k-fold.

Definition 2.5.1. *Holdout cross validation.* Prior to modelling, holdout CV partitions input data into two mutually exclusive subsets; training and test (holdout). For this method, a user-selected split is undertaken to (1) train the model and (2) test with unseen data. This aims to estimate how well the model performs on a new dataset. Holdout CV can sometimes provide pessimistic results because it only trains on a small proportion of data. Figure 2.9(a) provides a description of this method with a 60% training set and a 40% holdout set.

Definition 2.5.2. *k-fold cross validation.* KCV, on the other hand, partitions data into k subsets, performs the analysis on k-1 subsets (known as the training set) and validates the analysis on the remainder (the 'holdout' or 'test' set). The process is then repeated k times where the test set is a different subset each time. The validation results are then averaged across all folds to get a final result. Figure 2.9(b) provides a description of the KCV method with k=10.

Standard CV techniques share a common assumption; the training and test

set are independent of each other. This however is not appropriate for datasets containing autocorrelation, most notably; spatial and temporal data.

The prime complication that comes with data containing autocorrelation is the requirement to measure to what extent the unseen data is dependant upon the modelled data. For example, spatially autocorrelated data with second-order or intrinsic stationarity will be less correlated when the modelled data and the unseen data are further apart, and highly correlated when they are close together. Given that urban problems are inherently spatial in nature, Chapter 6 will discuss in great detail the challenges and potential solutions to dealing with dependencies between training and test sets.

2.5.1 Validation metrics

Model validation metrics provide quantitative measures to characterise the agreement between predictions and observations. In this thesis, I utilise three validation metrics: the squared Pearson correlation coefficient (r^2), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE).

Definition 2.5.3. r^2. The r^2 statistic measures a model's 'goodness of fit' to the real data and is defined in Equation 2.22.

$$r^2 = \left(\frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{(n\sum(x^2) - (\sum x)^2)(n\sum(y^2) - (\sum y)^2)}}\right)^2 \tag{2.22}$$

A perfect fit shows an r^2 of 1 and a poor fit has an r^2 of 0. This statistic provides a measure of how well the observed outcomes are replicated by the model. The output measures a relative value, showing whether predicted and observed data vary similarly i.e., at the same distance apart. The problem with this approach is that all model predictions can lay far apart from their observed counterparts but still perform perfectly because their distance apart is consistent. This could happen if the wrong model parameters are selected, i.e., nugget, sill or range. To confirm it's integrity, a second metric providing an absolute measure should be taken.

Definition 2.5.4. *RMSE*. The RMSE measures the differences between the modelled and observed values of random variables, such that:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{2.23}$$

The RMSE is the square root of the average of squared errors (residuals) and is therefore sensitive to outliers. The output RMSE value is always greater than or equal to 0 where an RMSE of 0 is a perfect fit. This metric provides an absolute understanding of how large the average error is. It should only be compared with other experiments using similar or identical data.

Definition 2.5.5. *MAPE*. The MAPE is a measure of difference between the predicted and observed data, represented as a percentage. In a scatter-plot, MAPE presents the average percentage distance along the y-axis between each modelled and observed point. MAPE is defined as:

$$MAPE = \frac{100}{n}(\sum_{i=1}^{n}\frac{y_i - \hat{y}_i}{y_i}). \tag{2.24}$$

MAPE is conceptually simpler and more interpretable than RMSE, most notably because MAPE does not require the use of any squares. However it is not defined where the actual value is 0, and it puts a heavier penalty for negative errors. Only the second of these shortfalls could affect the results of our thesis.

2.6 Urban Case Studies

This thesis considers two popularly discussed urban case studies: house prices and traffic flow. The purpose of this is threefold: (1) to examine the success of any **methodological** proposals stated in my contributions; (2) to provide an application used in **industry** and (3) to contribute to the **applied science of cities**. As such, the remainder of this section will review the current literature relating to both of these case studies, primarily urban house price prediction as

it features most heavily throughout the thesis.

2.6.1 House prices

The Office for National Statistics (ONS) is the most common supplier of house price predictions in the United Kingdom (UK). Their method finds the aggregate median price across a government defined 'output area' (more details published at [100]) for the whole of England and Wales. This model simplistically recognises the effects of proximity, neighbourhoods and spillover on house prices. The 'comparable sales' approach is another popularly employed method, particularly in industry, where the sale price of the (typically 3) closest properties are used to determine the value of a new property [40]. A final common method utilised in industry is 'spatial interaction' [49] which assesses a site by looking at its location from the consumer's perspective i.e., by identifying clusters of people [4] (target markets) based on their life stage and life style and then apply these clusters to a land's surrounding community for comparison.

Hedonic automated valuation models (AVMs)

The aforementioned models are typically small scale or non-granular. Large scale hedonic AVMs, on the other hand, are mathematical algorithms which exploit the availability of data to reliably understand the value of many real estate assets over a large area for a single point in time [37]. Hedonic valuation assumes that a heterogeneous product is a function of multiple attributes where each attribute has its own affective price on the good. This means that the sum of each attribute produces the final hedonic valuation [90].

Most contemporary AVMs are present in the machine learning domain [90, 99]. Such examples consider the effects of topography and natural geography [73], building footprints [102], school proximity [85], over head pylons [13] and crime [121] on house prices. Notably, it has been shown that space [37] followed by time [65] can infer up to 71% of a property's value, no other known variable can infer this much.

A framework central to the area of house price hedonic AVMs is put forth by [99] who hypothesises that a property's value (V) is some function of lot characteristics (L), structural characteristics (S), neighbourhood variables (N), accessibility variables (A), land use variables, proximity externalities (P) and data collection time (T). Although useful to inspire features that one may not have otherwise considered, it does not discuss the importance of appropriate feature selection to avoid the over fitting of certain statistical effects, for example, a hedonic model which assumes two highly correlated variables may overfit and not be suitable for generalisation to unseen data.

Typically, machine learning for house price prediction witnesses linear and multivariate regressions, RIPPER, C4.5, Adaboost and Case-Siller [16, 103]. All of these methods have produced indices of an aggregate output area such as cities, counties or regions. This decision is consistently taken with the argument that individual residential property sales are irregular and hence non-predictable in nature.

Spatial AVMs

Spatial models, on the other hand, typically attempt to look at all individual properties within a specific area. These methods are utilised in my thesis and assume that house prices can be predicted, irrelevant of the presence of previous sales data at the exact location of interest.

Contemporary literature recognises a considerable growth in utilising spatial technologies in real estate [40, 42, 86]. This is because structural, neighbourhood and accessibility characteristics are all a function of proximity [8, 14, 44]. In accepting this, researchers are commonly producing rent and price map surfaces [23], much like the one described in Figure 2.6. The most common methods in achieving such maps are Kriging and geographically weighted regression (GWR) [44, 50, 61, 75]. Examples include: mass appraisals [8] and spatial lag house price indices [113]. Interestingly, [104] did not build a continuous map, but instead undertook a similar model on lattice data.

A specific area of spatial AVMs for house prices is geostatistical modelling which was first motivated by [43] who, in his paper, argued that Kriging could replace the neighbourhood and accessibility variables typically used for hedonic house price prediction. His case study predicted the price of houses in locations where transaction data had not previously occurred in Baltimore. Thereafter [75] compared the results of several Kriging-based house price predictors obtaining a normalised RMSE of 1.019 in some cases; [60] noted that Bayesian maximum entropy can improve Ordinary Kriging, producing mean absolute errors as little as $7,000 on 2,700 homes in Texas and [20] introduced an Iterative Residual Kriging (IRK) method in Granada to present a MAPE of 16%. Each of these show that geostatistical modelling can accurately predict the price of houses. All of the aforementioned experiments utilise different data which make their methods hard to compare. Despite this, it seems consistently agreed that house price residuals are mostly related to space. I further prove this in Chapter 3 with a full analysis of a UK house price dataset.

Unlike some work, the aim of this thesis is not to maximise accuracy with the correct Kriging model, kernel or covariates, but instead to search for the optimal urban distance metric, which in turn supports kernel and hyperparameter selection.

2.6.2 Traffic flow

Traffic flow predictors are a subset of intelligent traffic systems (ITS). They are used to: assess potential designs for new road layouts; reduce accident hotspots and predict short-term traffic congestion [118, 131]. Temporal traffic predictions are most common and successful, utilising ARIMA [129], Markov chains [132], Bayesian Belief Networks (BBN) [118] and Artificial Neural Networks (ANN) [84]. These methods sometimes obtain mean absolute percentage errors optimised at 8.6% [126]. Although strong, all of these prior works make one key assumption, which may not be valid in several scenario's - data is present at all locations of interest.

Some applications i.e., road design or market analysis require information at locations with no sensors. This is where spatial analysis becomes useful. Given a set of sensors at location s_i, the total daily traffic flow Z_{s_i} has potential to infer the traffic flow for an unknown neighbouring location. The benefits of this is to minimise the number of sensors required to understand traffic flow fully. In fact, [38] puts forth that predictive analysis can reduce the computational overheads of feeding too many sensors for real time applications.

It is typical to use Kriging for the prediction of traffic flow, for example: [127] utilises Universal Kriging to predict the average daily traffic counts in Texas, obtaining a prediction error of 31%; [47] models Wake County, North Carolina using Ordinary Kriging; [69] models St. Louis in the State of Missouri and a novel spatiotemporal random effects model is implemented in Bellevue, WA boasting a MAPE of as little as 8% [130].

Finally, non-Euclidean spatial traffic flow models include a spatial moving average model to integrate the kernel against a white noise process [63]. By running the kernel upstream from a location, they rather sophisticatedly develop a valid flow model. Finally, an alternative, yet slightly less granular approach discussed the use of cost weighted distances between raster grids for urban traffic flow prediction [74].

2.7 The Practicalities of Storing, Retrieving and Analysing Spatial Data

Geographic Information Systems (GIS) facilitate many of today's spatial technologies by analysing and storing spatial data on a visual and interactive map. Spatial data can be inputted to a GIS in two types; *vector* and *raster*.

Definition 2.7.1. *Vector Data.* Vector data can be displayed as a point, polyline or polygon. Where a point is a precise location s in space, a polyline is a vertex connecting two or more points and a polygon is an ordered set of poly-

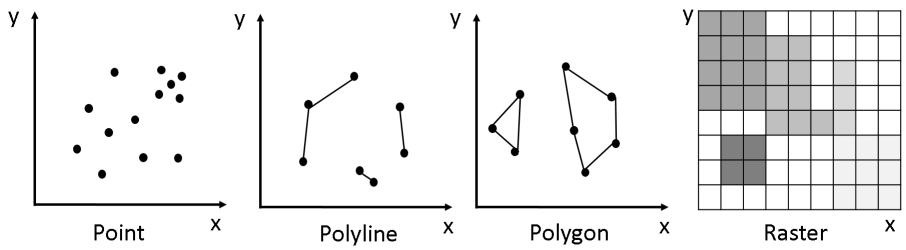


Figure 2.10: A visual comparison of all GIS data types: vector point, vector polyline, vector polygon and raster grid.

lines with a closed path. In addition, a feature class is a collection of features of the same type i.e., a set of house locations in one (.shp) file is called a point feature class and a set of house footprints in one (.shp) file is called a polygon feature class. Finally, each feature in vector data is also associated with a vector of attributes, for example the price of each house. The primary reason for selecting vector data over raster is it's geographical precision.

Definition 2.7.2. *Raster Data.* Raster data are (ir)regularly spaced grid cells, commonly described as pixels. Each pixel contains some attribute which determines the colour of the pixel in a GIS software. These types of data can appear pixelated if the attributes are not regular, close, continuous or spatially autocorrelated.

A satellite image is a typical example of raster data. Raster data types are usually smaller to store than vector data and quicker to process for calculations and plotting, however they are ordinarily based on some aggregate area, which may make these data types less precise compared to vector data.

Figure 2.10 provides a visual comparison between both data types, where the vector data (graphs 1-3) could be the location of 13 addresses, 3 road networks and 2 building footprints respectively. The raster data could be a satellite image showing the topography of an area. My research primarily utilises vector data for it's precision, granularity and accuracy, however Publication 3 does utilise a raster lidar dataset which shows the elevation of land above sea level.

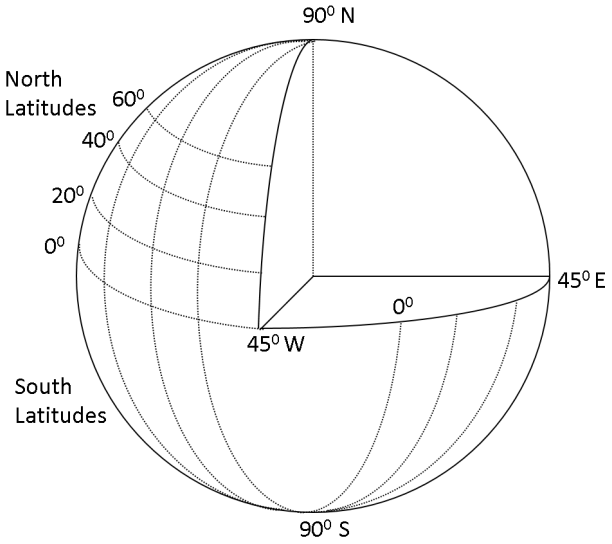


Figure 2.11: A visual representation of longitudes and latitudes on the earth's surface.

2.7.1 Referencing data

In order to analyse spatial data it must be referenced to a single geographic plane. As such, coordinate systems are structures used to reference spatial points to the earth's surface. Given that the earth is not a perfect sphere or ellipsoid and the fact that its surface is not smooth, precise locations can be hard to calculate, hence there are hundreds of localised coordinate systems, most notably, in the UK, the British National Grid. Additionally, the World Geodetic System (WGS) introduces the most accurate worldwide coordinate system. Within this thesis the data is all converted to WGS (commonly referenced as longitude (long) and latitude (lat)). With a consistent coordinate system, one can summarise distances, directions and paths between locations. Figure 2.11 provides the visual explanation of how the WGS coordinate system is calculated.

2.8 Final Remarks

The contents of this chapter has provided the reader with the necessary material required to fully understand the remainder of this thesis. Most notably, I have: (1) put forth a history of urban space theory; (2) introduced important spatial theory such as stochastic processes, data types, SAC and stationarity/heterogeneity; (3) extended discussion into the theory of geostatistical modelling; (4) discussed the different approaches to validating spatial models; (5) introduced a number of urban case studies and (6) described the practicalities of storing, retrieving and analysing spatial data.

CHAPTER 3

Datasets

In this chapter, I introduce two case studies utilised in my thesis; urban house prices and urban traffic flow. I also provide a description of all handled data. This chapter will be broken down into three sections: (1) distance data, (2) urban house price data and (3) urban traffic flow data.

3.1 Distance Data

A common theme throughout this thesis is the use of distance functions for geostatistical modelling and cross validation (CV). As such, this section describes how each distance is calculated or sourced. For geostatistical spatial modelling, each pairwise distance between points is represented as a distance matrix. For example, a distance matrix is a matrix where each row and column represent a single observation s_i and s_j respectively, the value where row i and column j meet is the distance between s_i and s_j (defined as $d_{i,j}$). Notably, spatial models are concerned with the inter-distance between all observations, hence the vector of rows s_i for $i \in 1 \dots n$ will be the same as the vector of columns s_j for $j \in 1 \dots n$. The pairwise distance matrix will hence have a 0 diagonal.

The pairwise distance matrices formed in this thesis will support six distances functions: (1) Minkowski, (2) Euclidean, (3) Manhatten, (4) road distance, (5) travel time and (6) combined road distance and travel time. Distance functions (1)-(3) are defined in Section 2.4.1 and are positive definite (PD) metrics. Distances (4)-(6) are defined fully in the following sections (Section 3.1.1 and Section 3.1.2).

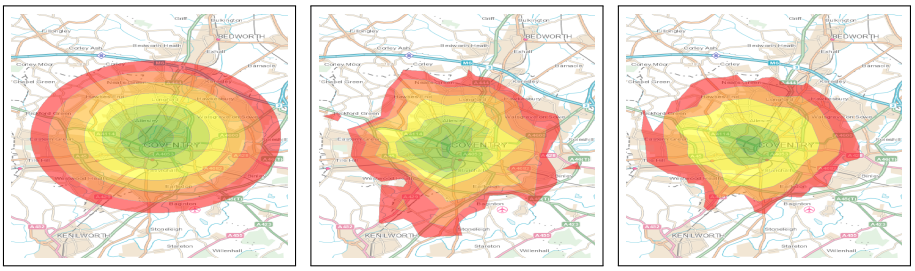
Figures 3.1(a)-3.1(c) show Euclidean, road and travel time distance isochrones. An isochrone is a line on a map connecting points relating to equal distance or

Figure 3.1: An isochronal comparison of Euclidean, road and travel time distances.

(a) Euclidean distance from 1 to 4 miles around the centre of Coventry City.

(b) Travel time distance from 1 to 10 minutes around the centre of Coventry City.

(c) Road distance from 1 to 4 miles around the centre of Coventry City.



times. Each isochrone is shown for 0 to 4 miles and 0 to 10 minutes of Coventry city centre. This figure emphasises the significant differences between distance measures, and hence the affect that they can have on a spatial model.

3.1.1 Road distance data

Within this thesis, there are two ways that the pairwise road distance matrix is calculated; *road network distance* and *restricted road distance*.

Road network distance data

The *road network distance* method is the current state-of-the-art for spatial modelling [136]. In order to get the pairwise distance matrix for all observations s_i in a dataset, one should first source the entire road network for their study area, which in my case is the United Kingdom (UK), sourced from Ordnance Survey (OS). This dataset is a polyline vector shapefile (as defined in Section 2.7). Figure 3.2 provides a small subset of the data in blue layered over a backdrop map sourced from OpenStreetMaps (OSM).

Given the entire road network and a set of observations, the *road network distance*:

1. Snaps the road network to each point. In my case, I search for the closest polyline in the OS UK road network against each observation then I

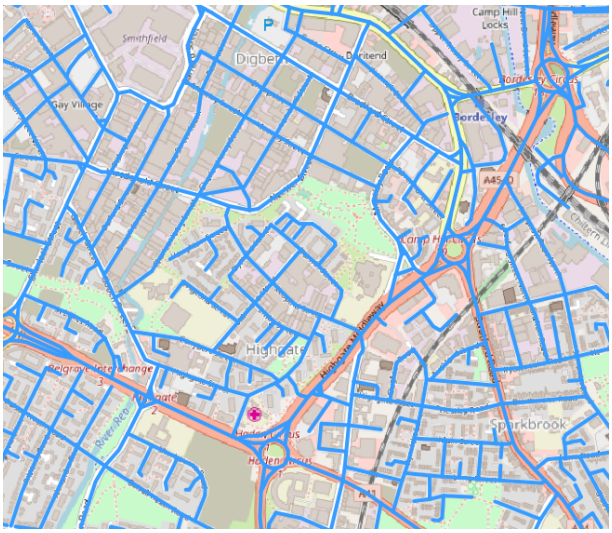
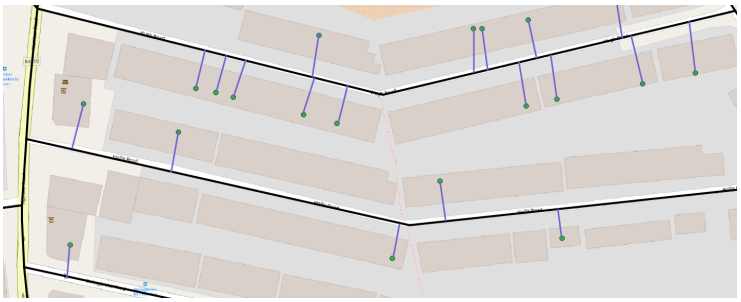


Figure 3.2: OS's road network dataset plotted on OSM's background street map.

add another polyline from that location to the observation. Figure 3.3(a) provides an example of this where the black lines are the actual road network, the blue lines are the snapped polylines and the green points are the observations. I then;

2. Find the shortest path between all pairwise distances. The method utilised for a 'road network distance' (in this thesis) is Floyd Warshall (FW), which is a commonly employed shortest path algorithm. The pseudo-code for Floyd Warshall can be found in Algorithm 1. Notably FW is one of the fastest shortest path algorithms and produces a PD pairwise distance matrix with a zero diagonal. These reasons and the fact that it is deemed the current state-of-the-art for spatial modelling is why I choose FW.



(a) Snap points to roads.

(b) Snap roads to points.

Figure 3.3: A visual description of the two methods for snapping observations and roads: snap points to roads versus snap roads to points.

Algorithm 1 Floyd Warshall.

Require: V, $w(u, v)$
Let: dist be a $|V|$ x $|V|$ array of minimum distances initialised to ∞
for each vertex v **do**
    dist[v][v] $\leftarrow$ 0
end for
for each edge (u,v) **do**
    dist[u][v] $\leftarrow$ w(u,v) (weight of the edge (u,v))
end for
for k from 1:$|V|$ **do**
    for i from 1:$|V|$ **do**
        for j from 1:$|V|$ **do**
            if dist[i][j] > dist[i][k] + dist[k][j]dist[i][j] $\leftarrow$ dist[i][k]+dist[k][j]
        end for
        END **if**
    end for
end for
Finish

A similar but alternative approach to step 1 would be to snap the points to their closest polyline i.e., move the point to the polyline. This approach is

visualised in Figure 3.3(b) where the green points are observations, the black lines are actual roads and the blue points are the snapped observations. This approach can move neighbours closer to each other. This is not beneficial for, say a house price case study, because large properties with large drive ways typically contain larger house price variations compared to those closer to the road [109].

Restricted road distance data

Alternatively, *restricted road distance* data is sourced from the Open Street Routing Machine (OSRM) and takes the shortest path along the OSM road network. The data is openly sourced from an API with a reasonable usage license.

This routing machine considers travel modes such as cars, bicycles and walking. The OSRM data also takes account for a number of road restrictions which are labelled by OSM. The most notable restrictions are one-way systems, path availability and speed limits. Table 3.1.1 provides just some examples of other restrictions considered. Throughout this thesis, 'restricted road distance' is utilised unless stated otherwise. The reason for undertaking a different approach will be for the purpose of comparison with other existing methods.

3.1.2 Restricted travel time distance data

OSM labels also consider *restricted travel time*; the time that is takes to travel between two points using a specific mode of transport. This routing option considers all of the same restrictions as road distance, and some more, including estimated congestion, speed limits, land gradients and the affect of certain road management systems such as traffic lights and pedestrian crossings. This provides potential for a newly sophisticated approach to measuring urban spatial autocorrelation (SAC).

Table 3.1: A subset of restrictions utilised in the OSRM's road network and travel time calculations from OSM labels.

Restriction type	Description
Mode	Car, walking, cycling, wheelchair access ...
Barrier	(Rising) bollard, cattle grid, border control, checkpoint, toll booth, sally port, (lift) gate ...
Restriction	Motor vehicle, vehicle, permissive, designated, destination, private, agricultural, forestry, emergency, parking aisle ...
Speed profile	Motorway, trunk, primary, secondary, tertiary, ferry, residential, living street, track, unclassified...
Surface speeds	Concrete, paved, cement, compacted, paving stones, metal, grass, gravel, unpaved, cobblestone, stone, sand, mud ...
Tracktype speeds	Grade 1-5, intermediate, bad, horrible, impassable ...
Maxspeed	Urban, rural, trunk, motorway, single/dual carriageway
U-Turn	Time in seconds
Traffic signal	Time in seconds
Oneway	Boolean, y/n
Route speed	Ferries, piers, movable bridges

3.1.3 Combined restricted road distance and travel time distance data

The last pairwise distance matrix considers a combination of the *restricted road distance* and *restricted travel time* distance matrices defined above. The purpose of calculating this is presented in Section 4.2 and Figure 4.1. I put forth two approaches to calculating this matrix:

1. The matrix is calculated with two user-defined weights (α_{rd} and α_{tt}). These weights are always positive and sum to 1. They are multiplied by the road distance and travel time matrices. This approach prioritises the matrix which is most influential for the application. The formula is $D_{comb}{=}\alpha_{rd}D_{rd} + \alpha_{tt}D_{tt}$ where D_{rd} and D_{tt} are the road and travel time pairwise distances. This method is utilised in Chapter 6 because it is highly practitioner motivated where application knowledge is assumed. Matrix A provides the format of the combined output distance matrix undertaken in this approach where $d^{rd}_{i,j}$ and $d^{tt}_{i,j}$ represent the road distance and travel time distance values between locations i and j.

Matrix A.: Combined matrix option 1.

$$\begin{bmatrix} \alpha_{rd}d^{rd}_{1,1} + \alpha_{tt}d^{tt}_{1,1} & \alpha_{rd}d^{rd}_{1,2} + \alpha_{tt}d^{tt}_{1,2} & \alpha_{rd}d^{rd}_{1,3} + \alpha_{tt}d^{tt}_{1,3} & \dots & \alpha_{rd}d^{rd}_{1,n} + \alpha_{tt}d^{tt}_{1,n} \\ \alpha_{rd}d^{rd}_{2,1} + \alpha_{tt}d^{tt}_{2,1} & \alpha_{rd}d^{rd}_{2,2} + \alpha_{tt}d^{tt}_{2,2} & \alpha_{rd}d^{rd}_{2,3} + \alpha_{tt}d^{tt}_{2,3} & \dots & \alpha_{rd}d^{rd}_{2,n} + \alpha_{tt}d^{tt}_{2,n} \\ \alpha_{rd}d^{rd}_{3,1} + \alpha_{tt}d^{tt}_{3,1} & \alpha_{rd}d^{rd}_{3,2} + \alpha_{tt}d^{tt}_{3,2} & \alpha_{rd}d^{rd}_{3,3} + \alpha_{tt}d^{tt}_{3,3} & \dots & \alpha_{rd}d^{rd}_{3,n} + \alpha_{tt}d^{tt}_{3,n} \\ & \dots & \dots & \dots & \dots \\ \alpha_{rd}d^{rd}_{n,1} + \alpha_{tt}d^{tt}_{n,1} & \alpha_{rd}d^{rd}_{n,2} + \alpha_{tt}d^{tt}_{n,2} & \alpha_{rd}d^{rd}_{n,3} + \alpha_{tt}d^{tt}_{n,3} & \dots & \alpha_{rd}d^{rd}_{n,n} + \alpha_{tt}d^{tt}_{n,n} \end{bmatrix}$$

2. The second option calculates a matrix using a linear regression of the road distance and travel time matrices. This is done such that the upper and lower triangle of both matrices are considered i.e., the regression has four features: upper triangle restricted road distance, upper triangle restricted travel time, lower triangle restricted road distance and lower triangle restricted travel time. This is because the distance matrices are not symmetric. This method is utilised in Chapters 4 and 5. Unlike method 1, this method assumes that the user has no knowledge of the application, this is beneficial for this thesis, which attempts to generalise across multiple urban applications. Matrices B and C are examples of upper and lower triangles.

Matrix B.: Example of an upper triangle.

$$\begin{bmatrix} d_{1,1} & d_{1,2} & d_{1,3} & d_{1,4} & d_{1,5} & \dots & d_{1,n-1} & d_{1,n} \\ 0 & d_{2,2} & d_{2,3} & d_{2,4} & d_{2,5} & \dots & d_{2,n-1} & d_{2,n} \\ 0 & 0 & d_{3,3} & d_{3,4} & d_{3,5} & \dots & d_{3,n-1} & d_{3,n} \\ 0 & 0 & 0 & d_{4,4} & d_{4,5} & \dots & d_{4,n-1} & d_{4,n} \\ 0 & 0 & 0 & 0 & d_{5,5} & \dots & d_{5,n-1} & d_{5,n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 0 & \dots & d_{n-1,n-1} & d_{n-1,n} \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & d_{n,n} \end{bmatrix}$$

Matrix C.: Example of an lower triangle.

$$\begin{bmatrix} d_{1,1} & d_{1,2} & d_{1,3} & d_{1,4} & d_{1,5} & \dots & d_{1,n-1} & d_{1,n} \\ 0 & d_{2,2} & d_{2,3} & d_{2,4} & d_{2,5} & \dots & d_{2,n-1} & d_{2,n} \\ 0 & 0 & d_{3,3} & d_{3,4} & d_{3,5} & \dots & d_{3,n-1} & d_{3,n} \\ 0 & 0 & 0 & d_{4,4} & d_{4,5} & \dots & d_{4,n-1} & d_{4,n} \\ 0 & 0 & 0 & 0 & d_{5,5} & \dots & d_{5,n-1} & d_{5,n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 0 & \dots & d_{n-1,n-1} & d_{n-1,n} \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & d_{n,n} \end{bmatrix}$$

These calculations, although simple, are effective in my results throughout, as such I have opened up a further avenue for research in the area of optimising this calculation - this is discussed further in Chapter 8.

3.1.4 Why travel time?

The intuition behind utilising travel time in addition to road distance is due to the fact that although road distance and travel time are correlated, some legal, customary and social restrictions are exclusive to travel time only; traffic flow, pedestrian crossings, road quality and so forth. These restrictions make travel time more accountable for human mobility patterns than road distance. In addition, practitioners can select travel time more dynamically (i.e., at different times of day) to better inform their own models.

Furthermore, different cities experience different road accessibility. For example, it may take 30 minutes to travel 1 mile in London, but only 2 minutes to travel 1 mile in Coventry; travel time takes this into account.

Finally, the combined road distance and travel time (RDTT) distance matrix affords the opportunity to take into account the exclusive behaviours of both matrices.

3.2 House Price Data

The 'house price' data is sourced from Her Majesty's Land Registry's (HMLR's) openly available 'Price Paid' database. This data is space and time stamped for

Table 3.2: Feature name, description and data type in HMLR's 'Price Paid' dataset.

Feature name	Description	Data type
UID	Unique transaction identifier	Integer
Property type	Flat, terraced, semi-detached, detached	String
Tenure	Freehold or leasehold	Binary
Date of transaction	Transaction date for the property	Date/Time
Address	Full address including postcode	String
Build status	Is the property newly built	Binary

all residential properties that have been sold in England and Wales since 1995. Table 3.2 provides a list of all the data available.

Throughout this thesis 'house prices' refer to the price that a property has sold for. This price may be different to the 'asking price' because sellers and purchasers typically negotiate. In addition, this thesis only considers the sale price and not the rental price, this is because these markets can be considerably different. Furthermore, 'house prices' do not refer to all 'residential properties' (i.e., any property sold for domestic use), it only refers to 'houses', i.e., the entirety of a detached, semi-detached or terraced property. Flats/apartments are not included throughout. The reason being that these markets act differently to each other due to the fact that 'houses' are sold as 'freehold' and flats are (typically) sold as 'leasehold'. A freehold purchase is one which includes both the sale of the property and the land it stands on, whereas a leasehold sale essentially rents the land of a property from the freeholder for a period of time i.e., the leaseholder does not own the land that their property lay on.

For this thesis, I only consider the freehold houses sold between the 01-January-2016 and the 01-January-2017. Chapters 4-6 utilise a spatial subset of Coventry, which accounts for 3,669 observations. The nationwide predictor discussed in Chapter 7 considers $\tilde{1}$15,000 observations.

The land registry's 'Price Paid' data provides an address, but no exact longitude and latitude locations. As such, I access the Ordnance Survey's (OS's) educationally available "Addressbase" dataset which contains all the address locations in the UK. This is done with a string match, pertaining to a 98% match

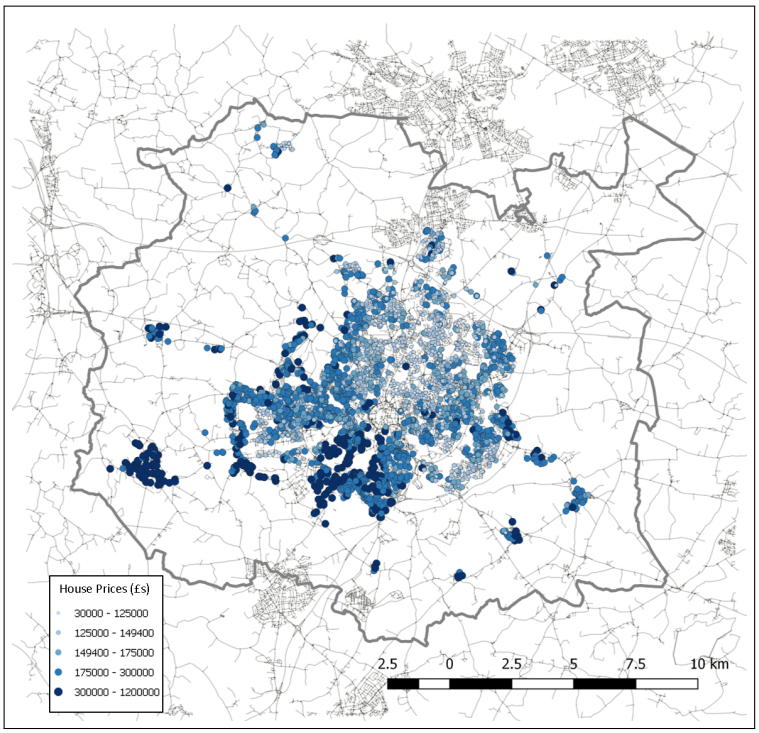


Figure 3.4: A visual description of the house price dataset across Coventry, projected to 2017.

rate. The remaining 2% are ignored in my experiments.

For Coventry, I undertake the standard Moran's I-test to confirm SAC. As expected, the houses dataset (containing 3,669 properties) show a strong result of $I_{\text{observed}} = 0.1559136 >> I_{\text{expected}} = -0.00267094$. In addition, a standard deviation of 0.001123158 and p-value $\rightarrow 0$ is measured. These results allow us to reject the null hypothesis that there is no SAC present at significance level $\alpha = 0.05$. These outputs emphasise the appropriateness of spatial interpolation with the Coventry house price data.

Finally, Figure 3.4 provides a visualisation of all observations considered in my experiments contained in Chapters 4-6. The larger and darker points represent the higher house prices. The smaller and lighter points embody smaller house prices. The black lines exhibit the road network in Coventry and the thick grey line serves as the border of the city.

3.3 Traffic Flow Data

All traffic flow data is openly sourced from the Highways Agency England's data portal. Throughout this thesis 'traffic flow' refers to the total daily count of traffic passing a specific sensor. Traffic considers all motorised vehicles and ignores foot traffic or push bicycles.

Across England, the number of sensors are 16,976, which is a small subset of the 139,019 miles of England roads [41]. My experimental analysis considers only 711 of these, which lay in the city of Birmingham. The observations (i.e., total number of motorised vehicles to pass the sensor in a day) considered in this thesis are the averages between 01 -January-2016 and the 01-January-2017'.

Table 3.3: Feature name, desctiption and data type in the traffic flow dataset.

Feature name	Description	Data type
UID	Unique transaction identifier	Integer
Collection method	Counted or estimated	Boolean
Road name	Name of the road	String
Location	Region, long and lat	string/numeric
Start and end junctions	Lookup to the junction dataset	Integer
Link length	Length of road between junctions	numeric
Small vehicles	Count of motorcycles and cars	numeric
Large vehicles	Count of buses, vans and lorries	numeric
All motor vehicles	Count of all motorised vehicles	numeric

Table 3.3 provides a list of all features from this dataset. Most notably, the longitude and latitude provides us with the ability to spatially model the data. Publication 2 utilises the remaining columns, however all of the work in our 3 contributions consider space only.

For this data, I conduct a standard Moran's I-test to confirm SAC. The results are $I_{observed} = 0.1604474 >> I_{expected} = 0.0002727025$ with a P-value of 0. These results allow us to reject the null hypothesis that there is no SAC present at alpha = 0.05. Hence, it is appropriate to engage in spatial interpolation with this data.

Figure 3.5 provides a visualisation of the points considered in my experiments. The larger and darker points represent the higher daily counts of mo-

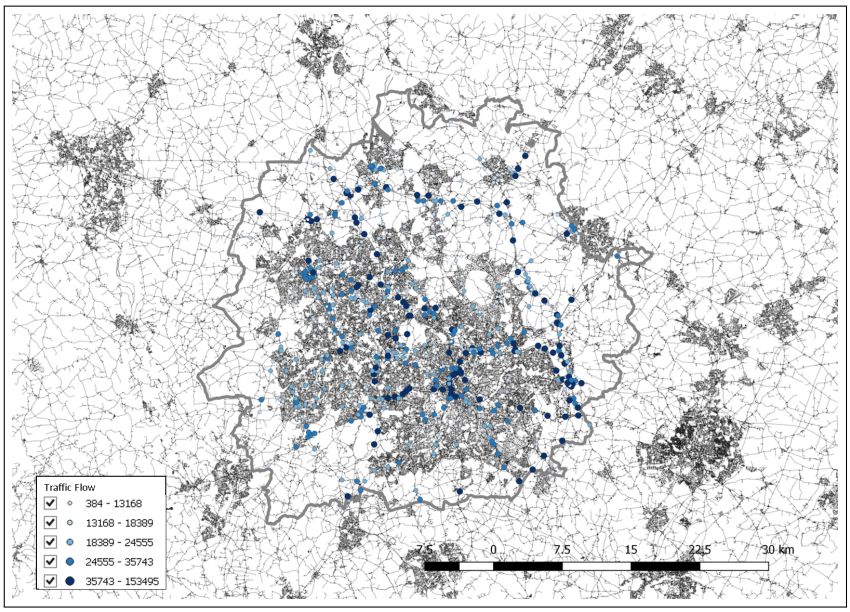


Figure 3.5: A visual description of the traffic flow dataset across Birmingham, projected to 2017.

torised vehicles. The smaller and lighter points embody the lower daily counts of motorised vehicles. The black lines exhibit the road network in Birmingham and the thick grey line serves as the border of the city.

CHAPTER 4

Modelling Space in the City; a Real Estate Case Study

In the following sections I describe the experiments undertaken to address **RQ1**. Each section contains the information required to populate my first of three key contribution chapters. The specific content from this chapter shows how I design an automated valuation model (AVM) to predict the price of residential property in Coventry, UK. I achieve this by means of geostatistical Kriging. Unlike traditional applications of distance based learning, this contribution is the implementation of non-Euclidean distance metrics by approximating road distance, travel time and a linear combination of both, which I hypothesise to be more related to urban house prices than straight-line (Euclidean) distance (rationale provided in Section 4.1). Given that, to undertake Kriging, a valid variogram must be produced (see section 2.3.1), my experiment exploits the conforming properties of the Minkowski distance function (with $P > 1$) to approximate a road distance and travel time metric. A least squares approach is put forth for variogram parameter selection and an Ordinary Kriging predictor is implemented for interpolation. The predictor is then validated with 10-fold cross validation (CV) and checkerboard hold out against the, almost exclusively employed, Euclidean metric. Given a comparison of results for each distance metric, one witnesses an r^2 of 0.6901±0.18 SD for real estate price prediction compared to the traditional (Euclidean) approach obtaining a suboptimal r^2 value of 0.66±0.21 SD. The results of this chapter are taken from Publication **4**.

4.1 Introduction

By 2030 investable real estate is expected to have grown by more than 55%; amounting to a UK residential market value of £9.145tn [93]. Consequently, leaders of real estate, policy makers and everyday home buyers are looking for information driven technological solutions to drive sustainable, low risk decisions in a newly global market [108]. In addition, the complex network structures, unprecedented urban growth and wealth of available real estate data makes (inter)national and urban residential markets more interesting and accessible than ever before. As such, machine learning algorithms, under the name of automated valuation models (AVMs), exploit this data to reliably understand the value of real estate over large areas where market behaviour may differ significantly. One such way to model these market behaviours is to utilise the vast data sources available to the single most influential variable which in this case is space (as seen in Section 4.2).

It was shown in Section 2.2.2 that spatial relationships require the removal of the assumption of independent and identically distributed (i.i.d) random variables for the purpose of predictive modelling. This is due to interdependencies between spatial points, known as SAC (as defined in Section 2.2.4). This is because an occurrence of dependency structures in spatial data introduces redundancy that must be taken into account to avoid an overestimation of statistical effects. As such, I put forth a spatial interpolation model named Kriging (see definition in Section 2.3.2) to predict house prices [88]. A prerequisite to Kriging is a variogram, which computes each pairwise distance h with a Euclidean function (definition 2.4.2). The Euclidean function is unrealistic for urban settings which contain complex physical restrictions and social structures for example road and path networks, large restricted areas of private land and legal road restrictions such as speed limits and one-way systems. This chapter hence hypothesises that the 'actual' space represented in the (Euclidean) semivariogram is currently ill-informed.

In this Chapter, I implement three new distance metrics into a set of house price Kriging predictors; (1) approximate road distance, (2) approximate travel time, and (3) a combination of both. My application puts forth a set of valid Minkowski distance metrics which are proven to better approximate restricted road distance and travel time across Coventry (UK) compared with a Euclidean distance.

4.1.1 Chapter structure

Section 4.2 provides a motivating example. Then, Section 4.3 reviews the existing literature related to this contribution. Thereafter, Section 4.4 describes the four-step method utilised within this chapter; collapsing time, distance matrix estimation, variogram fitting and spatial interpolation. Afterwards, Section 4.5 validates the method with a case study of 3,669 real-estate transactions across Coventry. Finally, Section 4.6 concludes the chapter with some discussions regarding further work opened up by this research.

4.1.2 Contributions

The key primary contributions within this chapter are: (1) a Minkowski approximation of a pairwise restricted road distance metric utilising OpenStreetMaps (OSM) data; (2) a Minkowski approximation of a pairwise travel time metric utilising OSM data; (3) a Minkowski approximation of a pairwise combined restricted road distance and travel time metric utilising OSM data; (4) a comparison study of house price predictors in Coventry with distance metrics (1)-(3) against a commonly used Euclidean metric. The final contribution shows that spatial interpolation can be improved with non-Euclidean city-motivated distance functions.

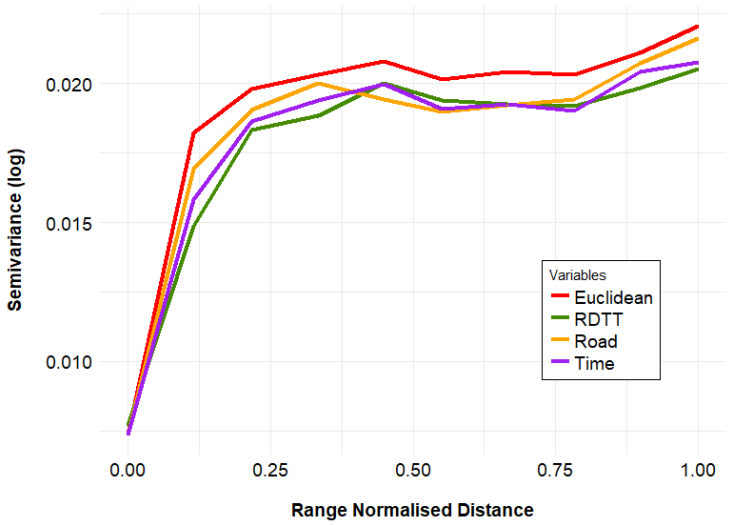


Figure 4.1: A comparison of variances for urban house prices with different distances; Euclidean, road distance ("Road"), journey time ("Time") and a linear combination of road distance and journey time ("RDTT").

4.2 Motivating Example

Figure 4.1 provides an analytical proof of my hypothesis. The figure shows four variograms with my house price case study; Euclidean, road distance, travel time and combined (RDTT). The combined matrix simply applies a 50% weighting to both road distance and travel time. The road distance and travel time estimates are normalised and symmetric. This semivariogram is crudely built using a range normalised distance and an upper triangle symmetry. It can be seen that the non-Euclidean distance metrics contain lower semivariances for the price of pairwise urban residential properties compared to Euclidean. This result provides motivation to undertake a more sophisticated estimation of road distance, travel time and a combination of both, which is undertaken in this chapter.

4.3 Background Reading

4.3.1 House prices in space

Since the early 19th century, space and distance have been theorised as the primary functions for property valuation. For example; favoured prices are given to those properties within close proximity to its central market place [123], community centre [70] or central business district [15]. Most contemporary analysis mimics this trend, for example predicting property value by using (1) the average sales price of other properties in the local comparables market, (2) a spatial clustering of properties and demographics [86] and (3) a local demographic 'trade area' analysis [40]. More detail of these methods are discussed in Chapter 2.

Most contemporary machine learning based AVMs are hedonic in nature (a function of multiple attributes) [90, 99]. Attributes relating to residential property price include; topography and natural geography [73], building footprint [102], school proximity [85], over head pylons [13] and crime [121].

In addition, [102] describes the implementation of a spatiotemporal autoregressive model on 70,822 properties in Fairfax county from 1961 to 1991. Their prediction, with twelve variables, reduced the median absolute error by 37.35% relative to an indicator-based model. Additionally, [37] used a house price Kriging predictor to produce an r^2 of 0.72 on a nationwide United Kingdom (UK) AVM. Finally, [65] put forth a geographically and temporally weighted regression (GTWR) for house price prediction, in which an r^2 of 0.88 is achieved on a dataset of residential house sales in Calgary (Canada) between 2002 and 2004. Each of these approaches highlight the importance of space to house price prediction. Although, these methods consider space and proximity to be represented by a Euclidean distance only.

4.3.2 Non-Euclidean distance based predictors

Manhatten [51, 122], Geodetic [6] and water-based (shortest path over water) [97] distances have all been implemented in distance based learning algorithms,

each showing some minor improvements compared with the Euclidean function. All of these methods are motivated by some access-restricted environment; city-based routing, world distances and smooth edges respectively. In addition, I hypothesises that road distance and travel time are intrinsic to contemporary house price modelling, and it is these features that I approximate in this chapter.

Without direct access to the above datasets, it cannot be confirmed that the input metrics (Manhatten, geodetic or water-based) produce a valid variogram. As such, [39] discusses dimensionality reduction to approximate a Euclidean metric from a (potentially invalid) non-Euclidean metric input. Using simulated data with isotropic spatial dependence, their work builds four omnidirectional variogram estimators, showing that their newly defined "Stream" distances consistently outperform the standard Euclidean function, whilst always remaining valid.

Similarly, [136] produces a Kriging predictor with a road distance network using an Isomap algorithm; a variation of isometric embedding. The predictor estimates traffic flow in Nanchang, China. This method uses the Floyd Warshall algorithm to build a non-restricted road network. This does not consider accessibility restrictions such as one way systems or traffic lights.

With regards to the use of Minkowski distances for spatial modelling, [115] approximates the distance between a set of postcodes and a hospital with a 1×N vector of Minkowski distances. The selected Minkowski p-value was the one which was most correlated with the shortest path along the Calgary road network. The results from their paper motivates the experiment in this contribution, however I uniquely introduce an N×N distance matrix with a Minkowski p-value most correlated to travel time, restricted road distances and a combination of both.

Finally, a Minkowski distance metric is also put forth with geographically weighted regression (GWR) [82]. Their work tests GWR with a combination of Minkowski p-values (1-8, inf) at intervals of 0.25. Their paper puts forward the interesting point that, for each dataset a new p-value may need to be cal-

Algorithm 2 Collapsing time - distance matrix selection - variogram parameter selection - spatial interpolation.

Require: K_{ord}, d^p, D, maximum likelihood estimator (MLE)
Input: $D = \{\mathbf{X}_t^s, \mathbf{Y}_t^s, \}_{\substack{s=\{1:S\} \\ t=\{t_0:\Delta t:T\}}}$
Temporal mapping to time τ:
$D^\tau \leftarrow g(D) \ \ \forall \ \ t, s \in \{t_0 : \Delta t : T\}, \{1 : S\}$
Stratified sampling: Sample across each LSOA
$D_\sigma^\tau \sim \sigma_{\text{stratified}}(D^\tau)$
for z in {road distance, travel time, linear combination} **do**
    $d_z^p = \text{argmax}_p \, r^2(d^p, z)$
end for
for z in {road distance, travel time, linear combination} **do**
    for V in 10-folds, Checkerboard **do**
        Variogram selection on $vs \leftarrow MLE(Train(D_\sigma^\tau), d_z^p)$
        Ordinary Kriging on $Prices \leftarrow K_{\text{ord}}(Train(D_\sigma^\tau), Test(D_\sigma^\tau), vs)$
return r^2, RMSE, MAPE
    end for
end for
Finish

culated, which can be time-consuming on large datasets. Notably, both GWR and Kriging are local-spatial prediction models, however Kriging prediction is regularly noted as an improvement to GWR [89, 91].

4.4 Scientific Method

In this section I describe the experiment undertaken; the prediction of Coventry house prices. There are 4 stages of this experiment, each described in Algorithm 2:

1. Collapsing time: converting a discrete, non-uniform, spatiotemporal sold price dataset D into a uniform time singular sold price output D^T utilising a space-time cube comparison.

2. Distance matrix estimation: calculating a set of Minkowski coefficients to predict road distance and travel time between all house price points.

3. Spatial prediction: Ordinary Kriging on a sample of 3,669 house price observations (data described in Section 3.2).

4. CV and validation metrics: model validation against five distance metrics (Euclidean, Manhatten, road distance, travel time and a combination of both) using 10-fold CV and checkerboard hold out.

4.4.1 Stage 1: collapsing time

The 'Price Paid' dataset (herein named D) described in Chapter 3 is utilised. This accounts for 3,669 sales in Coventry. Stage 1 predicts each property's sale price based on its value on the 01-January-2017 (for time singularity). This process involves each property being assigned some percentage price change based on the date that it was sold and the lower super output area (LSOA) that the property is contained within to produce a value for all 3,669 properties at the date 01-January-2017 (D^T). The errors, for the purposes of this experiment, are minimal or non-existent due to the small temporal and granular spatial areas being considered. Section 3.2 provides more details about how this dataset is sourced.

4.4.2 Stage 2: distance matrix estimation

Consider a one way system in a city road network, where one route may be longer than it's counterpart route. Hence, a restricted road distance matrix containing such a restriction will not be symmetric. Figure 4.2 shows an example where the distance between houses A to B is 0.24 miles along the red dotted line which takes a route along 'Brownshill Green Road'. This road is marked as a one way system, this means that the route B to A must be different (0.44 miles). The same reasoning applies for a travel time distance matrix.

Given my extensive discussion on distance metrics and variogram validity in Chapter 2, one must ensure that the road distance and travel time functions are metric. The example above shows that a road distance and travel time matrix does not always satisfy P3 and P4, hence a metric estimate is required. A simple method of making the distance matrix symmetric would be to (1) duplicate the lower triangle, (2) select the minimum or maximum of the lower/upper triangle

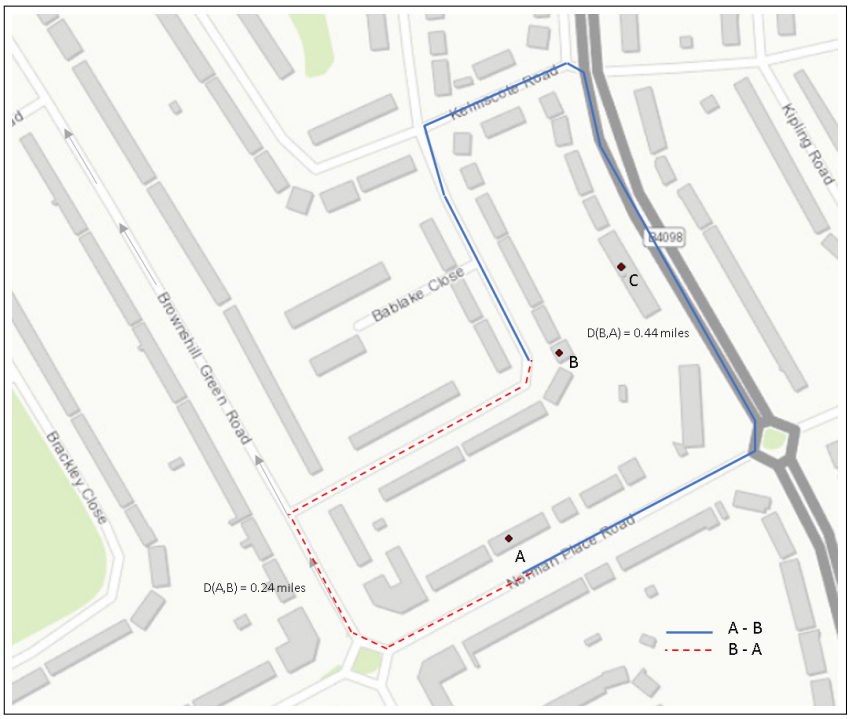


Figure 4.2: A visual example of where P3 and P4 are not satisfied.

or (3) calculate the average between route $A \rightarrow B$ and $B \rightarrow A$. However, this doesn't always overcome the problem of P4 as a shorter non restricted route could potentially be found. The experiment in this chapter instead considers Minkowski coefficients (definition in Section 2.4.1). Assuming that there is a Minkowski p-value that is similar to road distance, travel time or a combination of both, then this Minkowski value can be used as a valid estimate of road distance and/or travel time.

Estimation optimisation

For this experiment, three scenarios are attempted:

1. The Minkowski p-value with the highest r^2 value to the *restricted road distance* matrix described in Section 3.1.1. For my dataset, one finds that it is p=1.55.

2. The Minkowski p-value with the highest r^2 value to the *restricted travel*

time matrix described in Section 3.1.1. For my dataset, one finds that it is p=1.7.

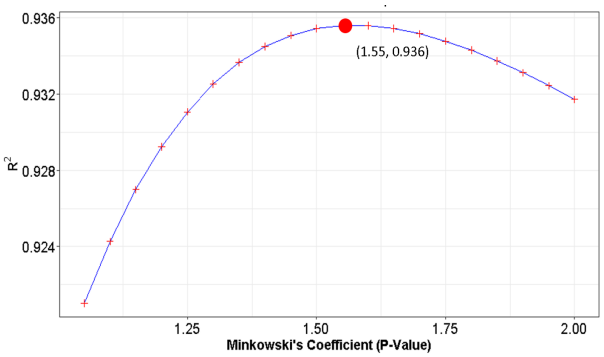
3. The Minkowski p-value with the highest r^2 value to the *combined road distance and travel time* matrix described in Section 3.1.1. For my dataset, one finds that it is p=1.6.

Figure 4.3 shows the r^2 value for each Minkowski p at 0.05 intervals between 1 (Manhatten) and 2 (Euclidean). It can be seen that a combination of the two distance matrices has the highest r^2 (=0.946) at $p = 1.6$, which shows that Minkowski coefficients are strong at predicting a restricted road network compared with Euclidean or Manhatten distance matrices.

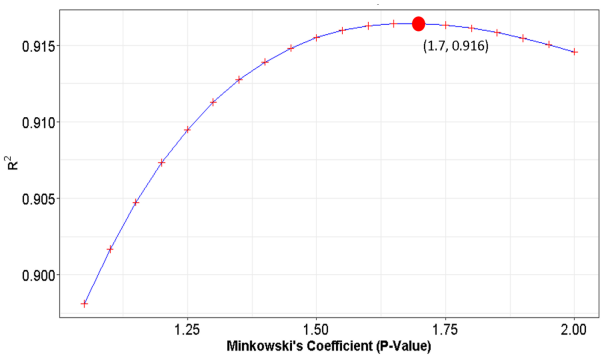
The combined road distance and travel time matrix is calculated as a linear model with four variables; (1) Road distance A→B; (2) Road distance B→A; (3) Travel time A→B; (4) Travel time B→A, as described in Section 3.1.3. This is an approach which to my knowledge has never before been undertaken and attempts to fully understand the spatial utility function (defined in Section 2.1) of a house purchaser. Figure 4.4 provides a comparison of each physical distance (Euclidean, Manhatten, actual road (in both directions) and Minkowski p= 1.6) between two points.

4.4.3 Stage 3: variogram fitting and spatial interpolation

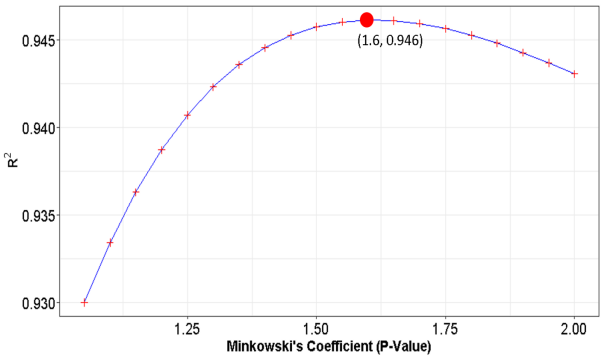
For comparison, I run a separate experiment for all of the selected Minkowski p-values ($p = [1.55, 1.6, 1.7]$) i.e., those most correlated with road distance, travel time and a combination of both. I also run the experiments with the current state-of-the-art; Manhatten and Euclidean distances. It is not surprising to find that the same model (Matern) is optimal in all cases because each distance matrix used to plot the lag are highly correlated to each other, as seen in Figure 4.3. The (hyper)parameters are selected using maximum likelihood estimates (MLE) for random fields. It are these hyperparameters that affect the output predictions, see Section 2.3.1 for details on each hyperparameter. In this chap-



(a) OSRM road distance versus Minkowski p-value goodness of fit graph.



(b) OSRM travel time versus Minkowski p-value goodness of fit graph.



(c) OSRM linear combination of road distance and travel time versus Minkowski p-value goodness of fit graph.

Figure 4.3: The 'goodness of fit' value for each Minkowski coefficient, tested against the OSRM's actual road distance calculations, travel time calculations and a linear model of both.

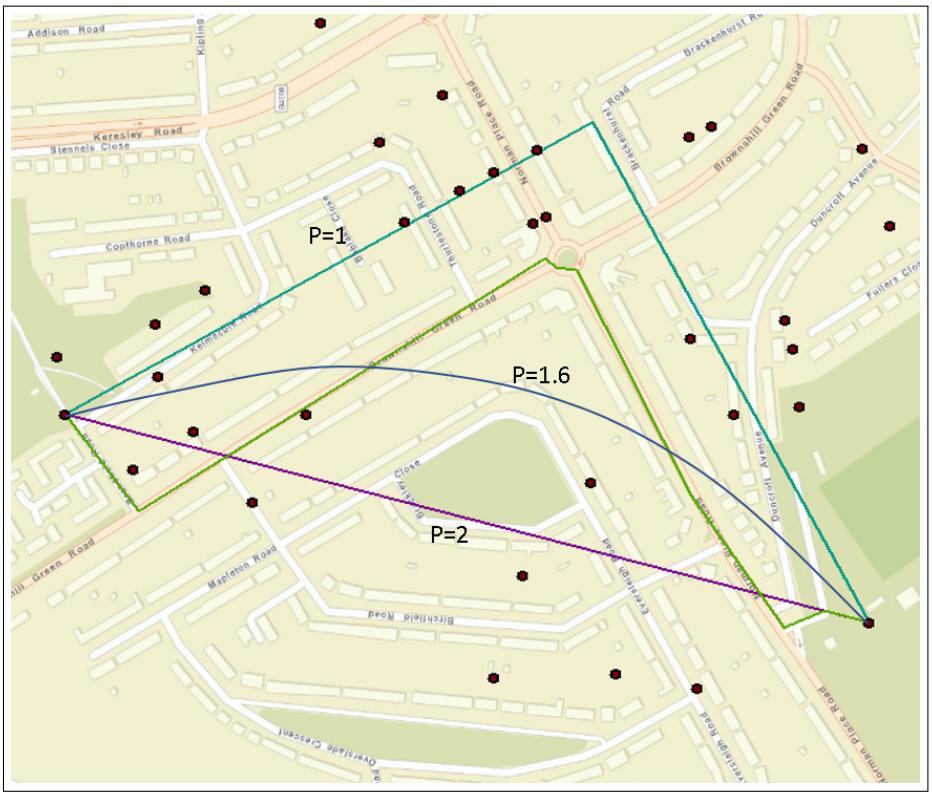


Figure 4.4: A streetmap comparing distance functions; road, Euclidean, Manhatten and Minkowski distance.

ter, I consider Ordinary Kriging only, as defined in Section 2.3.2. The reason for this is because I am only interested in optimising prediction with appropriate distances functions rather than Kriging models. This provides a potential further avenue of research.

4.4.4 Stage 4: cross validation and validation metrics

k-fold cross validation (KCV) and checkerboard holdout, as defined in Section 2.5 are the CV methods utilised in this Chapter. The checkerboard method attempts to take into account over fitting as a result of spatial dependency between the training and test set. This method provides a training sample of 1,832 properties and test sample of 1,837 properties. Figure 4.5 shows the checkerboard polygons used to separate the training and validation test set

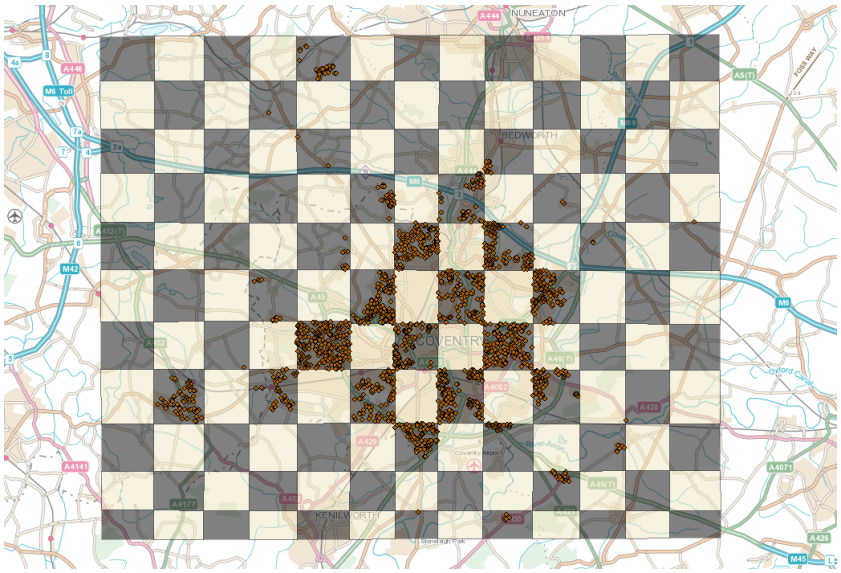


Figure 4.5: Spatially aware checkerboard sampling polygons utilised for my hold out method.

for each experiment. Chapter 6 argues that this is not the best approach for validating spatial data because the removal of observations from the training set can cause a pessimistic estimation of generalisation performance.

The experiment's success is measured on a number of validation metrics: (1) the squared Pearson correlation coefficient (r^2); (2) Root Mean Squared Error (RMSE) and (3) Mean Absolute Percentage Error (MAPE). A paired T-test is also undertaken to state whether the results are statistically significant enough for the null hypothesis that the price of a house can be predicted by space only. Each of these metrics and statistical tests are defined in Section 2.5.1.

4.5 Results

In this chapter, I present an approach to estimate restricted road, restricted travel time and combined distances using the Minkowski distance function. I then undertake five spatial interpolations, each containing the same observation data and different Minkowski distance metrics, using Ordinary Kriging. In this

Table 4.1: Results for 10-fold cross validation.

	10-fold cross validation				
Distance matrix	p=1 (Manhatten)	p=1.55	p=1.6	p=1.7	p=2 (Euclidean)
r^2	0.683	0.6847	0.6901	0.6843	0.663
RMSE	57,115	57,000	57,013	57,439	58,913
MAPE	17.92%	17.9%	17.895%	18.01%	18.12%

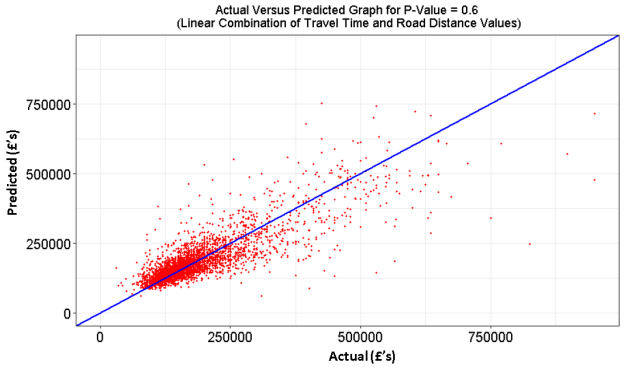
Table 4.2: Results for checkerboard holdout.

	Checkerboard stratified validation				
Distance matrix	p=1 (Manhatten)	p=1.55	p=1.6	p=1.7	p=2 (Euclidean)
r^2	0.4509	0.4514	0.4558	0.4499	0.4418
$RMSE$	82,414	82,367	81,940	82,507	82,972
$MAPE$	24.52%	24.51%	24.40%	24.53 %	24.57%

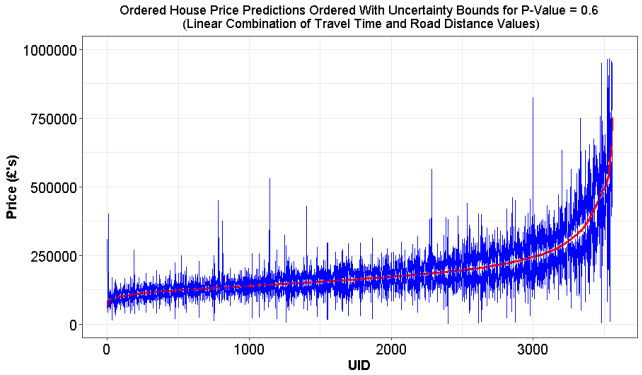
Section, I provide a set of results for each experiment.

Tables 4.1 and 4.2 provide the validation results for each model showing that non-Euclidean distance metrics can produce a more appropriate set of parameters for house price prediction in Coventry. Notably, the Minkowski metric which is most related to a combination of restricted road and travel time distances has the best performing interpolation with an r^2 (0.6901), supporting my hypothesis that urban house prices are more related to road distance and travel time than the popularly employed Euclidean and Manhatten metrics (Minkowski of 2 and 1 respectively). Figure 4.6(a) visualises the prediction versus actual price for all properties trained with my best performing distance matrix (p=1.6). In addition, Figure 4.6(b) shows the uncertainty bounds between folds for all properties in the 'Price Paid' dataset. The t-value and p-value of the best performing model are 1.312 and 0.1896 respectively, showing that space as a single variate is weak on its own; some more covariates could really support the model.

In general, the results show that residential valuation contains some spatial autocorrelation (SAC) which, with the use of appropriate distance metrics, can be improved. In addition, a student's t-test between experiments is calculated to show that the best performing (p=1.6) and poorest performing (Euclidean) Kriging outputs provide a statistically significant change with a p-value



(a) Actual versus predicted graph for p=1.6.



(b) House price prediction graph with uncertainty bound for p=1.6.

Figure 4.6: Results graphs for the best performing experiment.

of 0.0458. This is an appropriate test because the two populations have very similar (almost equal) variances (86,555 and 86,657 respectively). If this were not the case, I would have considered a Welch t-test [135].

4.6 Final Remarks

In this chapter, I have (i) converted a discrete, non-uniform, spatiotemporal sold price dataset D into a uniform time singular sold price output D^T utilising a space-time comparable process in Coventry; (ii) deployed a novel method of producing N×N road distance and travel time predictions; (iii) produced a novel N×N combined road distance and travel time matrix; (iv) calculated 5

variograms, each with a different distance function; (v) produced a spatially aware Ordinary Kriging calculation to predict house prices. Each of the models are tested using MAPE, RMSE and adjusted r^2. The optimal experiment with a combined road distance and travel time approximation yielded an adjusted r^2 value of 0.69 compared with the traditional Euclidean approach at 0.66.

Future work is to include: (1) testing the hypothesis with other applications and spatial interpolation methods; (2) implementing the findings into the SPENT algorithm from Publication 4 and (3) introducing a set of covariates to improve the overall accuracy of the model.

In the following chapters, I will introduce: (1) a new approach to better estimate valid distance functions from invalid matrices such as road distance and travel time for spatial modelling and (2) a state-of-the-art urban spatial CV method for estimating the generalisation performance of spatial predictive models along the range of interpolation to extrapolation scenario's. Thereafter, Chapter 7 puts forth a set of answers to the research questions (RQ) posed in Section 1.1, matches those answers with my results, and discusses the implications of my work to urban science, geostatistics and real estate. Finally, Chapter 8 concludes all of my findings and puts forth a set of research avenues that are opened up by this thesis.

CHAPTER 5

Producing a Valid Urban Spatial Model with Road and Travel Time Distance Functions

Urban environments are restricted by various physical, regulatory and customary barriers such as buildings, one-way systems and pedestrian crossings. These features create challenges for predictive modelling in urban space, as most proximity-based models rely on Euclidean (straight line) distance metrics which, given restrictions within the urban landscape, do not fully capture any spatial urban processes. In this Chapter, I continue to argue that road distance and travel time provide an effective measure of city mobility and hence I develop a new low-dimensional Euclidean distance metric based on road distance and travel time using an isomap approach. This method intends to improve the results displayed in Chapter 1 and open the research area to further urban problems above and beyond real estate.

This chapter's primary methodological contribution is the derivation of two symmetric dissimilarity matrices (B^+ and B^{2+}), with which it is possible to compute low-dimensional Euclidean metrics for the production of a positive definite (PD) covariance matrix with commonly utilised kernels and non-valid, non-Euclidean input spaces. This new method is implemented into a Kriging predictor and is used to estimate house prices of 3,669 properties in Coventry, United Kingdom (UK). I find that a metric estimating a combination of road distance and travel time, in both $\mathbb{R}^2$ and $\mathbb{R}^3$, produces a superior house price predictor compared with alternative state-of-the-art methods, that is, a standard Euclidean metric in $\mathbb{R}^{\mathbb{N}}$ and a non-restricted road distance metric in $\mathbb{R}^2$ and $\mathbb{R}^3$. Finally, I undertake an extensive comparison of cross validation (CV) techniques and I select the best model for predicting house prices in new locations, based on

the model's estimated generalisation performance on unseen data. This chapter addresses **RQ2** fully, and the results are taken from Publication **5**.

5.1 Introduction

By 2030, it is expected that 5 billion people will live in urban areas, 662 cities will have at least 1 million residents and there will be a total urban spread of 1.2 million km^2 [12, 98, 114]. Hence, cities will continue to accommodate over 50% of the world's population. In the UK, over 82% of citizens live in its 64 cities, a figure which has grown by more than 13% in the past 30 years [18]. Many UK cities suffer from legacy infrastructure, such challenges are well documented: Housing supply is not matching demand [62]; Commuting times are increasing [54] and there are shortages in services for the most vulnerable citizens [116]. Issues of urban growth and sustainability motivate the development of mathematical tools and models for explanatory and predictive analysis [124].

Urban models provide insight into the relationship between some chosen target value, house prices for example, and other potentially related variables, such as topography [73], building footprints [102] and crime [121]. Space [37] and time [65] consistently feature in most urban models, for example in house price prediction [35], traffic flow prediction [136] and in the analysis of green space and its impact on well-being [64]. A typical approach to understanding spatial characteristics in this way is through geostatistical proximity-based modelling. An example of this approach is Kriging (defined in Section 2.3.2), which assumes random variables to be spatially dependent and non-stationary over space. A common assumption in geostatistical models (including Kriging) is that proximity is based on Euclidean distance; this is in spite of the fact that dispersion in a city landscape is unlikely to exhibit such properties.

Traditionally, research in real-estate price modelling has considered distance to a specific location (e.g., workplace) and/or comparable prices of other submarkets within close proximity. A more sophisticated approach to this is to

include physical barriers such as buildings, road layout and non-accessible open space to the models, as distance, in practice, is clearly governed by such obstacles. This is evident in recent work on road-distance-based Kriging, which has been shown to be highly effective for urban house price prediction [35].

This chapter presents a natural extension to this earlier work by including travel time. In so doing it integrates a number of otherwise difficult to capture variables such as traffic flow, road layout, junction priority and congestion caused by on-road parking. The primary purpose is to show the effect that road distance and travel time have on predictive modelling; note I do not prescribe reasons for these effects (i.e., I will not be considering any covariates).

The methodological advances are, again, motivated by my work in urban house price prediction; that is, I attempt to model unexplained variation through proximity between observations, to underpin and improve on hedonic pricing models already available in academia and in industry.

As discussed in chapter 4, an essential prerequisite to geostatistical models is the production of a variogram and covariance function. Covariance and variogram functions must remain valid - PD and conditionally negative definite (CND), respectively [39], (see Section 5.4.1 for formal definition).

Given the extensive geostatistical research using Euclidean pairwise distances, there is no guarantee that any non-Euclidean distance matrix (PD or otherwise) will produce valid covariance or variogram functions. For this reason, pairwise road distance and travel time matrices are unlikely to be valid. Hence, the purpose of this research is to propose an isometric embedding approach with which one can approximate road distance and travel time in a lower-dimensional Euclidean space, to allow physical properties of cities to be represented in spatial prediction whilst still producing mathematically valid approximations.

In order to illustrate the benefits of these new distance metrics, I again, utilise my Coventry house price data to build a real estate automated valuation model (AVM). This AVM is used to provide mathematically modelled individual market values for 3,669 properties. The case study in Section 5.5 shows that

a combination of road distance and travel time produces a superior Kriging predictor compared with a Euclidean approach for all assessed validation metrics with my data.

5.1.1 Contributions

The contributions within this chapter, and above and beyond my previous chapter are as follows:

- First, methodological contributions are made via the derivation of two symmetric dissimilarity matrices (B^+ and B^{2+}), with which it is possible to compute low-dimensional Euclidean metrics for the production of a PD covariance matrix with commonly utilised kernels and non-valid, non-Euclidean, input spaces;
- Second, I demonstrate the application of this new geostatistical approach to the calculation of (i) approximate restricted road distance, (ii) approximate travel time and (iii) combined road distance and travel time matrices, in each case within an embedded lower-dimensional Euclidean space;
- Third, I compare a number of the most popularly employed CV techniques to assess the ability of each to estimate how well my model generalises to unseen data.

5.1.2 Chapter structure

The remainder of this chapter is organised as follows: background research is detailed in Section 5.2; Section 5.3 motivates the need for this research through two practical examples; new methodological contributions are described in Section 5.4 and applications of these methods, to urban house price prediction, can be found in Section 5.5, utilising the data introduced in Section 3.2. The chapter concludes in Section 5.6 in which I also document avenues for future research.

5.2 Related Literature and Key Concepts

5.2.1 Constructing optimal urban Kriging predictors

Kriging is a geostatistical spatial predictor which accounts for spatial covariance. The method utilises observation distances to understand the spatial structure of a dataset and hence determine its own interpolation parameters [33]. Kriging is used extensively for interpolation by ecologists [81], geographers [19] and geoscientists [67]. A full description of this method is put forth in Section 2.3.2. Notably, parameter optimisation, kernel selection and lag sizes are the primary strategies used in optimising experimental variograms and Kriging algorithms [30, 53, 133].

Kriging is commonly used in urban science and examples of its application include traffic flow prediction [136], travel time prediction [94] and trip planning [80]. The use of Kriging for urban real estate pricing is motivated by [8, 37, 44] who together note that space and time are highly influential in house price prediction. Each of these approaches however use Euclidean distance only and are discussed in detail in Section 2.6.

A small number of non-Euclidean distance-based approaches have been employed to Kriging, including those based on Minkowski (see Chapter 4 and [51, 35, 122]), geodetic [6] and water-based (shortest path over water) [97] distances. Each offers its own benefits, however it is difficult to assess whether each produces valid experimental variograms without access to the initial data; I show in Section 5.3 that relying on the fact that input distances are PD metrics is no guarantee of valid variograms.

Research which bears similarity to my own can be found in [83], who use geographically weighted regression (GWR) and a non-Euclidean distance metric for predicting London house prices. Their research also utilises road distance and travel time, however is limited to network shape and speed limit; my measures include a wealth of other data provided by OpenStreetMaps (OSM), see Table 3.1.1.

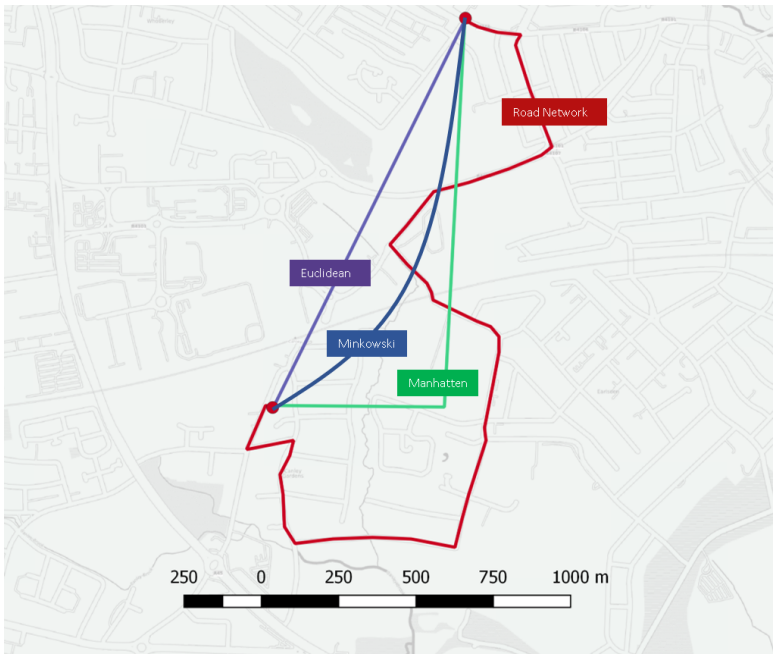


Figure 5.1: A comparison of the actual road, Euclidean, Minkowski and Manhatten distances between two points on a map [101].

Approaches based on GWR have advantages, in particular because there is no requirement for the matrix to be Euclidean (the matrix w_i of weights is diagonal, hence there is no need to check for positive definiteness, which is not the case with the covariance matrix used in Kriging [39]). However, it is noted in [35] (and Chapter 4) that Kriging typically outperforms GWR in spatial pricing models; this is especially true when implemented locally, which is the case in Ordinary Kriging which assumes intrinsic stationarity (i.e., a moving mean but a stationary variance between any two points).

5.2.2 Overcoming non-metric input spaces

For the most part, geostatistics relies on the assumption that each set of distances lie in a metric space $(\mathbf{M}, d)$, as defined in Section 2.4.1. There are three known methods which ensure that a distance matrix is valid (that is, that it produces a PD covariance matrix): The first uses isometric embedding to ensure a Euclidean input; the second is the use of kernel convolution, so that the kernel

fits any matrix and the third is to select a matrix which produces a valid covariance matrix. Previous research has assumed that the distance matrix must satisfy P1-P4 to produce a valid spatial interpolation. I do not subscribe to this view, as the example in Section 5.3 highlights.

With regard to the three methods that ensure matrix validity: Isometric embedding provides a dimensionality reduction technique with which it is possible to build a low dimensional Euclidean approximation of non-Euclidean inputs for variogram modelling. Using simulated data with isotropic spatial dependence, [39] builds four omnidirectional experimental variograms, each representing an α norm, for $\alpha = 1,\ldots,4$ ($\alpha = 2$ is Euclidean). When this data is applied with Kriging, the newly defined 'stream' distances outperform Euclidean distances in all cases; this is therefore my method of choice.

I note that other research proposes similar approaches to approximate road distance metrics, see [120, 136]. In [136] the Floyd Warshall algorithm is applied to a road network to estimate the actual road distance between pairwise locations. I note however that Floyd-Warshall only selects the shortest distance, irrespective of restrictions such as transport patterns and one way systems.

The use of kernel convolutions, which can be used to express moving averages, assume that correlated data can be expressed as linear combinations of uncorrelated data. This method has been successfully applied by [29], however I note that this method can be difficult to implement on problems with large datasets and is hence not considered further in this work.

Finally, the selection or creation of a valid covariance function can be undertaken. For example, [39] noted that a set of Manhatten distances produced non-valid variograms with Gaussian, Matern and spherical kernels, but were valid for an exponential kernel. I am aware that this approach has several restrictions and is also time consuming to compute, and so for this reason isometric embedding remains my method of choice.

5.3 Motivating Examples

The contributions within this chapter are based on the following two assertions: (1) that road distance, travel time and a combination of both are better indicators of urban proximity than the commonly utilised Euclidean function, and (2) that the only way to guarantee that a covariance matrix and variogram function are valid in this context is to ensure that a Euclidean distance metric is input for their calculation. Chapter 4 provides evidence of the first claim and Subsection 5.3.1 explores the second assertion also.

5.3.1 Calculating a valid variogram

To ensure that a variogram is valid, the input must be Euclidean. This implies that even PD distance functions cannot always produce a valid variogram, a concept which has potential to invalidate much previous research.

Non-PD inputs

Non-PD matrices produce non-PD kernels (covariance functions) which is usually as a consequence of the L_2 norm; note the next subsection provides other examples where this is the case. Matrices 1 and 2 below show a set of possible pairwise distances. These matrices are not symmetric, much like a road network containing one-way systems, and hence they are not PD. To test whether each matrix always produces a valid variogram, I select a Gaussian covariance function ($C(h) = \sigma^2 e^{(-h/a)^2}$) with σ^2=0.5, 0.08 and a=450, 1.5. The output vectors from this calculation are shown in Vectors 1 and 2 below.

Matrix 1.: Road Distance (Meter)

	1	**2**	**3**	**4**	**5**	**6**	**7**
1	0	266.5	459.4	738.1	602.5	614.3	640.6
2	266.5	0	321.6	600.3	464.8	476.5	502.8
3	459.4	321.6	0	278.7	143.1	154.9	181.2
4	738.1	600.3	278.7	0	346.6	358.4	342.4
5	602.5	464.8	143.1	346.6	0	358.4	342.4
6	614.3	476.5	154.9	358.4	222.8	0	133.8
7	640.6	502.8	181.2	384.7	249.1	133.8	0

Matrix 2.: Travel Time (Minute)

	1	**2**	**3**	**4**	**5**	**6**	**7**
1	0	0.81	1.188	1.186	1.71	1.628	1.75
2	0.702	0	0.855	1.523	1.38	1.29	1.42
3	1.133	0.8	0	0.67	0.522	0.44	0.56
4	1.8	1.47	0.67	0	0.96	0.982	1.05
5	1.55	1.212	0.412	0.956	0	0.603	0.723
6	1.681	1.348	0.548	0.98	0.72	0	0.44
7	1.7	1.36	0.56	0.99	0.72	0.44	0

Vector 1.: Road Distance and **Vector 2.: Travel Time**

$$\begin{bmatrix} 2.09991 \\ 0.74078 \\ 0.27006 \\ 0.22365 \\ 0.13790 \\ 0.04218 \\ -0.0145 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0.38814 \\ 0.098924 \\ 0.03598 \\ 0.018321 \\ 0.010134 + 0.00194i \\ 0.010134 - 0.00194i \\ -0.014469 \end{bmatrix}$$

In view of the negative roots in Vectors 1 and 2, it is clear that both covariance functions are not CND ($\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j C(h) \geq 0$) and hence road distance and travel time are not valid for variogram modelling.

PD inputs

Additionally, non-Euclidean PD matrices may also produce non-PD kernels, a fact that some previous research has been known to overlook. Matrix 3 below represents the same roads as in Matrix 1 and 2 above, but this time, the road distance is not restricted (much like the work by [136]); that is to say, one-way systems are not considered and hence are completely PD. The same covariance function and hyperparameters are used.

Matrix 3.: Road Distance (Meter)

	1	**2**	**3**	**4**	**5**	**6**	**7**
1	0	266.5	459.4	738.1	602.5	614.3	640.6
2	266.5	0	321.6	600.3	464.8	476.5	502.8
3	459.4	321.6	0	278.7	143.1	154.9	181.2
4	738.1	600.3	278.7	0	346.6	358.4	384.7
5	602.5	464.8	143.1	346.6	0	222.8	249.1
6	614.3	476.5	154.9	358.4	222.8	0	133.8
7	640.6	502.8	181.2	384.7	249.1	133.8	0

Vector 3.: PD Road Distances

$$\begin{bmatrix} 2.1346 \\ 0.74503 \\ 0.30465 \\ 0.153779 \\ 0.12919 \\ 0.039961 \\ -0.0072856 \end{bmatrix}$$

Vector 3 shows that the output eigenvector still contains negative roots, which itself means that the covariance function is not CND, despite the input matrix being PD.

This motivates my new approach for estimating non-Euclidean, non-PD distance matrices in a Euclidean space in order to produce valid covariance and variogram functions.

5.4 Method

I now describe how current state-of-the-art approaches estimate city-based proximity (i.e., non-Euclidean distance metrics). I also compare these approaches to my new method. I show how my proposed approach, isometric embedding with newly defined symmetric dissimilarity matrices (B^{+} and B^{2+}), produces a PD covariance matrix. As a result of this, I then show application of this new technique to the establishment of an urban real estate price predictor.

5.4.1 Distance matrix calculation

To undertake geostatistical modelling, a pairwise distance metric is required. This pairwise distance metric is populated with distances $d_{i,j}$ from a list of locations $\{x_i, i = 1, \ldots, n\}$ in Euclidean space $\mathbb{R}^n$. The matrix provides the basis for a valid metric if all $d_{i,j}$ satisfy P1-P4, see Section 2.4.1.

As I have previously shown, road distance and travel time are not natural metrics. Given this, I compare four methods for calculating conforming geostatistical distance metrics from these inputs: a Euclidean distance; a Minkowski approximation of restricted road distance and travel time; an isomap estimate of road distance and a newly formulated improved isometric embedding approach to estimating restricted road distance and travel time. Each distance is allocated a subsection below.

Euclidean distance

Unless otherwise stated, it is typical to assume a Euclidean function when referring to distance. Assuming two sites as vectors $\mathbf{s}=(s_1, \ldots, s_d)^T$ and $\mathbf{u}=(u_1, \ldots, u_d)^T$ in Euclidean space $\mathbb{R}^d$, then the Euclidean distance is defined in Section 2.4.1, where d is the number of dimensions (or attributes) and s_i and u_i are attributes.

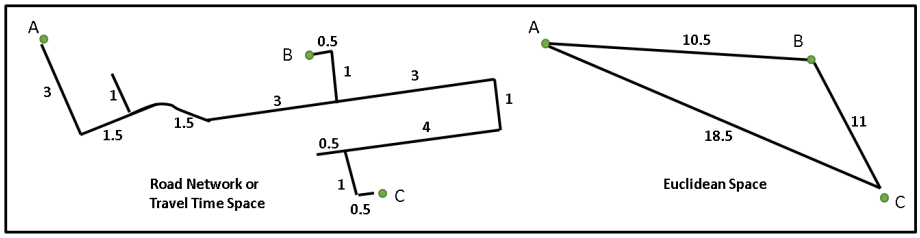


Figure 5.2: Illustration of the spatial transformation from road distance (or travel time) into a Euclidean space.

Minkowski distance

Assuming the same notation as above, the Minkowski distance is also defined in Section 2.4.1, where P is a user defined parameter. Manhatten and Euclidean distances are special cases of Minkowski, with P={1,2} respectively. In [35] and Chapter 4, I show that Minkowski distances with $P \neq \{1,2\}$ can better estimate road distance and travel time compared with Manhatten or Euclidean distances.

Isometric embedding and isomap

Isometric embedding provides the spatial transformation of a new metric space $\zeta'=(s', d')$ from $\zeta = (s, d)$, with point set $s = (s_1, s_2, \ldots, s_n)$, distance function $\mathbf{D}$ of ζ and distance function $\mathbf{D}'$ of ζ'. All associated s and d_{ij} values are intrinsic. If $\mathbf{D} \simeq \mathbf{D}'$ then the transformation still preserves topological adjacency among points in the original space ζ. Dimensionality reduction is a good means of achieving isometric embedding; multidimensional scaling (MDS) is the most popular such scheme.

Isomap, in addition to isometric embedding, attempts to detect the intrinsic characteristics of non-linear data, in which ζ may be a non-metric space. For example, isometric embedding assumes a Euclidean distance, whereas isomap supports other spatial features such as non-restricted approximate road [136] and geodesic [6] distances on a set of discrete points [120]. Figure 5.2 provides an example of a road distance layout (left) transformed into a low-dimensional Euclidean space (right) using isomap.

As stated, MDS is a dimensionality reduction technique used to achieve isometric embedding or isomap. Given an input metric **D** (which is for example Euclidean) in n-dimensional metric space ζ, the first stage of MDS is to calculate the dissimilarity matrix B

$$B = \frac{1}{2}\{a_{ij} - a_{i\cdot} - a_{j\cdot} + a_{\cdot\cdot}\} \tag{5.1}$$

where $a_{i\cdot}$ is the average of all a_{ij} across j. Formally, each element B_{ij} in matrix B is calculated by:

$$B^*_{ij} = \frac{1}{2}\Big(- d^2_{ij} + \frac{1}{n}\sum_{l=1}^{n} d^2_{il} + \frac{1}{n}\sum_{l=1}^{n} d^2_{lj} - \frac{1}{n^2}\sum_{l=1}^{n}\sum_{m=1}^{n} d^2_{lm}\Big) \tag{5.2}$$

where **B** is a new set of isometric distances which mimics a kernel where **B** is doubly centered. Although **B** is semi-PD, it is not guaranteed to produce a PD covariance function or a CND variogram (see proof in Section 5.3). **B** is definitely valid only when the input distance matrix **D**=$\{d_{ij}\}_{nxn}$ is Euclidean. Given this, classical MDS requires that the eigenvalues of B are $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_\alpha$, where α is a user-selected value based on an optimal κ:

$$\kappa = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} |\lambda_i|}, \tag{5.3}$$

and $\lambda_\alpha > 0$. The optimal κ provides the smallest value of α given some user-defined minimum variation threshold. Thereafter, the corresponding eigenvectors ($\Gamma = \epsilon_i$, for $i = 1, \ldots, \alpha$) are calculated. The penultimate step of MDS is to calculate a new dataset of points in the new α-dimensional subspace ζ'=(s', d'), where $s' = \Gamma\Lambda^{\frac{1}{2}}$ and $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_k)$. This new s' point set is the isometric subspace which best describes point set **D**; this process is called eigenvalue decomposition and explains the variance of the data in a lower dimension. In the final stage of isomap, the new coordinates in s' are used to calculate a new approximate distance metric using the Euclidean function.

If some inputs are non-metric, such as may be the case with travel time or

restricted road distance, the dissimilarity matrix $\mathbf{B}$ may not be semi-positive definite with an L_2-norm, a property which is essential for MDS. For this reason, a new B^+ dissimilarity matrix is proposed in which $\mathbf{D}$ is forced to be symmetric within the calculation:

$$B_{ij}^{+} = \frac{1}{2}(-\frac{1}{2}(d_{ij}^2 - d_{ji}^2) + \frac{1}{2n}(\sum_{l=1}^{n} d_{il}^2 + \sum_{l=1}^{n} d_{jl}^2 + \sum_{l=1}^{n} d_{lj}^2 + \sum_{l=1}^{n} d_{li}^2) - \frac{1}{n^2}\sum_{l=1}^{n}\sum_{m=1}^{n} d_{lm}^2). \tag{5.4}$$

Additionally, B_{ij}^{2+} takes a combination of both road distance and travel time matrices (the maximum and minimum distances are normalised between 0 and 1) to produce isometric distances, where δ_{ij} represents the normalised road distance and τ_{ij} represents the normalised travel time distance between each i and j:

$$\begin{aligned} B_{ij}^{2+} = \frac{1}{2}(-\frac{1}{2}(\delta_{ij}^2 + \tau_{ij}^2 - \delta_{ji}^2 - \tau_{ji}^2) + \frac{1}{2n}(\sum_{l=1}^{n}(\delta_{il}^2 + \tau_{il}^2) + \sum_{l=1}^{n}(\delta_{jl}^2 + \tau_{jl}^2) \\ + \sum_{l=1}^{n}(\delta_{lj}^2 + \tau_{lj}^2) + \sum_{l=1}^{n}(\delta_{li}^2 + \tau_{li}^2)) - \frac{1}{n^2}(\sum_{l=1}^{n}\sum_{m=1}^{n}\delta_{lm}^2 + \tau_{lm}^2)) \end{aligned} \tag{5.5}$$

Each new B_{ij}^{+} and B_{ij}^{2+} solves the problem of non-symmetry for travel time and restricted road networks, or indeed any non-PD matrix. This ensures that B is semi-PD, so that the process of MDS and the output distance matrices are also both valid. B_{ij}^{+} and B_{ij}^{2+} are the key contributions of this chapter.

'Stress' validates the effectiveness of classical MDS - it tests the goodness of fit for $\mathbf{D}'$ with the input metric $\mathbf{D}$ (the normalised sum of squares), such that:

$$Stress = \sqrt{\frac{\sum_i \sum_j (d_{ij} - d'_{ij})}{\sum_i \sum_j d_{ij}^2}}. \tag{5.6}$$

However, when implementing non-metric inputs, Stress should be calculated differently such that $d_{ij}^{b^+}$ and $d_{ij}^{b^{2+}}$ are the Euclidean functions on space B^+ and B^{2+} respectively. The reason for this is because I am no longer reconstructing elements d_{ij}. Rather, I reconstruct the dissimilarity matrix for the new metric space. A metric space can be confirmed such that:

$d_{ij}^2 = (\overrightarrow{b_i} - \overrightarrow{b_j})^T(\overrightarrow{b_i} - \overrightarrow{b_j})$ where $(\overrightarrow{b_i} - \overrightarrow{b_j}) = [b_{i1} - b_{j1}, \ldots, b_{in} - b_{jn}]$

hence

$d_{ij}^2 = (b_{i1} - b_{j1})^2 + (b_{i2} - b_{j2})^2 + \ldots = \sum_{d=1}^{n}(b_{in} - b_{jn})^2$ (Euclidean).

Given that one can define a Euclidean metric from B, one can be assured that it is indeed a valid metric space.

5.5 Case Study

Real estate valuation has become a much more data-driven and quantitative process. This said, the process of estimating the value of a property or land parcel through market appraisal remains the de rigueur of skilled market professionals. Having now worked in this domain for several years, my aim has been to scale-up and semi-automate the use of big data for real estate valuation.

To this end I build a so-called AVM for a sample of 3,669 residential properties in the city of Coventry in the UK, using Ordinary Kriging with a target valuation date of 1 January 2017. I develop a new approximate road distance and travel time metric for variogram calculations. For the purpose of comparison, and to ensure robust results, I run six experiments where each contain a different input distance metric:

1. Euclidean (vector norm of 2);
2. Optimal Minkowski (P=1.6)[35] ;
3. Floyd Warshall on a road network (PD road) [136];
4. OSM road distance with restrictions;
5. OSM travel time with restrictions;

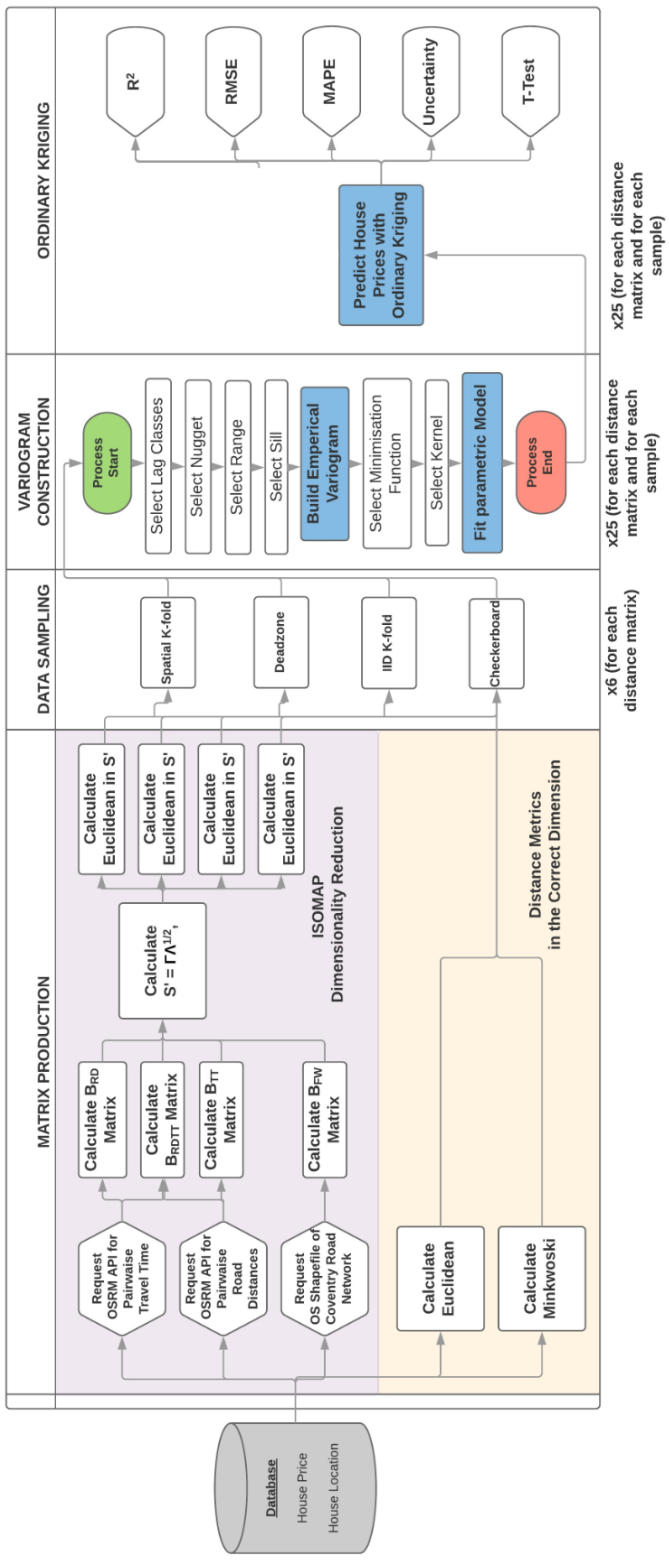


Figure 5.3: A flow diagram depicting the entire experimental process for the UK real estate valuation case study described in this chapter.

Algorithm 3 Pseudocode for the entire isomap algorithm displayed in the purple coloured section of Figure 5.3.

Require: $D = \{d_{ij}\}$, $\zeta = (s, d)_i$, x, Floyd Warshall, B^{+}_{ij}, B^{*}_{ij}, B^{2+}_{ij}, κ, S.

for Experiment in 3 to 6 **do**
  Let: $\zeta = (s, d)_i$ be a metric space with point set $\mathbf{s} = (s_1, s_2, \ldots, s_n)_i$ and distance function d and $\mathbf{x} = x_n$ is the point set of midpoints for each vertex.
  if {i=3}
  $D = \{d_{ij}\} \leftarrow$ Floyd Warshall
  Map D to a semi-PD distance metric with Eq. (5.2)
  elseif {i=4,5}
  $D = \{d_{ij}\} \leftarrow$ OSRM restricted road distance, travel time
  Map D to a semi-PD distance metric with Eq. (5.4) $(r < n)$
  else
  $D = \{d_{ij}\} \leftarrow$ OSRM restricted road distance, travel time
  Map D to a semi-PD distance metric with Eq. (5.5) $(r < n)$
  end if
  Embed into low-dimensional Euclidean space $\zeta' = (s', d')_i$ such that $\alpha < r$ (Eq. 5.3)
  Collect new coordinates s' given S' in ζ'
  Calculate the new Euclidean distances
end for

6. A combination of normalised road distance and travel time with restrictions.

Each experiment is subsequently referred to using the numerical identifier (1-6).

5.5.1 Data description

Our AVM uses input data regarding all houses that were sold in Coventry in 2016. For each of these 3,669 properties, the percentage change in house price, between the date sold and 01-01-2017, is calculated using the predicted change in value in each output area as defined by the UK Office for National Statistics. This provides a predicted price per property for the data 1 January 2017.

The datasets that I use are all open source and have been obtained from Her Majesty's Land Registry and the Ordnance Survey respectively. In addition, experiment (3) requires road network data, which is also sourced from the Ordnance Survey. Experiments (4)-(6) all require distances between points along a roadway and the time that it takes to travel these distances, this is sourced from the Open Street Routing Machine (OSRM) powered by OSM. All data is

described fully in Chapter 3.

5.5.2 Matrix construction

I process five distance matrices for the six experiments. Experiments (1)-(2) require a Euclidean and a Minkowski distance metric respectively, which are valid for variogram modelling (see the beige-coloured portion of Figure 5.3). Experiment (2) uses a P-value of 1.6 which was previously reported to perform best on the same dataset, see [35]. Experiments (3)-(6) require preprocessing using isomap (see purple-coloured portion of Figure 5.3). Experiment (3) utilises a road network to calculate a shortest path using the Floyd Warshall (FW) algorithm. Experiment (3) embeds the input distance matrix using dissimilarity matrix B^*. Experiments (4) and (5) embed the distance matrices sourced from OSRM and dissimilarity B^+. Finally, experiment (6) utilises the same distance matrices sourced from OSRM but now implementing the B^{2+} dissimilarity matrix. This entire process is depicted in the Matrix Production column in Figure 5.3, the purple-coloured portion of which is captured in Algorithm 1 and used in experiments (3)-(6).

Table 5.1 shows how successfully each calculated metric represents OSRM's road distance and travel time using a matrix goodness of fit value r^2. It can be seen that my embedded metrics with restrictions (experiments (4)-(6)) are best at approximating the actual distances. This means that, assuming my hypothesis that house prices are related to their pairwise proximity along a restricted road network (measured by travel time), I expect experiments (4)-(6) to outperform the other experiments in terms of my final Kriging predictor.

5.5.3 Data sampling for cross validation

The most sophisticated validation sampling techniques (hold-out and k-fold) assume data in both the test and training sets to be independent of each other. This is an assumption that may be unrealistic with datasets containing spatial autocorrelation (SAC), especially if the purpose of modelling is for interpolation

Table 5.1: The r^2 values for each distance metric compared with actual road distance and travel time matrices.

Experiment	Distance Metric	Actual Road Distance (r^2)	Actual Travel Time (r^2)
1	D_{Euc}	0.377	0.359
2	D_{Mink}	0.379	0.359
3	D'_{FW}	0.374	0.365
4	D'_{RD}	0.621	0.592
5	D'_{TT}	0.606	0.614
6	D'_{RDTT}	0.446	0.419

or close proximity extrapolation [107]. As such, four sampling techniques are considered, three of which consider spatial dependence for comparison (see 'Data Sampling' in Figure 5.3):

1. 10-fold cross validation on the full dataset of 3,669 properties;
2. Spatially stratified 10-fold cross validation (SS-KCV) on the full dataset of 3,669 properties;
3. Checkerboard holdout on a training set of 1,832 properties, with a test set of 1,837 properties;
4. Spatial k-fold cross Validation (spatial k-fold cross validation (S-KCV)) [107] on samples of the entire dataset, with each sample including 3,187 properties ± 135 for each fold [107].

K-fold cross validation (KCV)

K-fold CV randomly partitions a dataset into k equally sized subsets. One of these subsets is retained for testing, whereas the other k-1 are considered for training. For each fold, a different subset is retained for testing until all k subsets are tested. Figures 5.4(e)-5.4(f) show two of the ten folds in my six experiments. K-fold CV overestimates statistical effects on spatial random variables and hence produces an optimistic estimate of generalisation performance for unseen data.

Checkerboard holdout

Checkerboard holdout trains approximately 50% of the data and tests the remaining data based on whether they lay in the black or white grid squares (see Figure 5.4(a)). The case study uses a training and test set of 1,832 and 1,837 properties respectively. Checkerboard holdout is quick to apply, simple and removes some SAC. On the other hand, it removes a significant amount of training data and still contains bias at block borders.

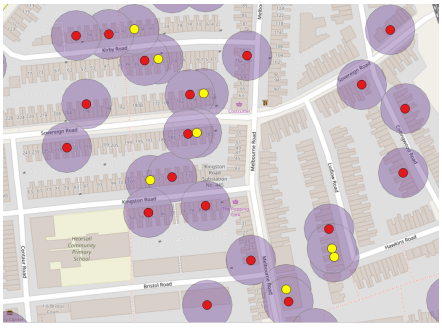
Spatially stratified *k*-fold cross validation (SS-KCV)

SS-KCV processes data in a similar manner to standard k-fold, however the data splits are spatial and not random. Two of the ten folds are shown in Figures 5.4(c)-5.4(d). As can be seen, each test subset is spatially separated from the training set, which can appropriately remove some bias caused by SAC. However, the data splits still contain SAC at and near sample borders.

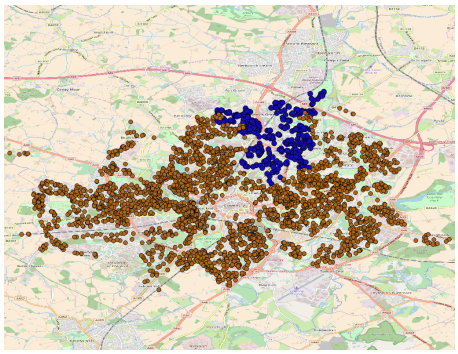
Spatial *k*-fold (S-KCV)

S-KCV estimates a predictor's performance by implementing traditional k-fold cross validation (KCV), whilst at the same time removing all training points within an empirically designed Euclidean dead-zone from all test points [107]. Figure 5.4(b) demonstrates this method where training points within 20 meters of each test point are in yellow for a specific fold. This method more efficiently removes SAC than the other methods. However, it relies on a user-defined dead-zone with no given heuristic and removes training points which in turn can cause pessimistic results. For the case study, I apply 20-metre zoning, which removes approximately 8% of the total training points: This parameter value is selected as it is at this level that I see the most significant change in results; close inspection shows that this removes on average 3 to 5 of a properties' closest neighbours.

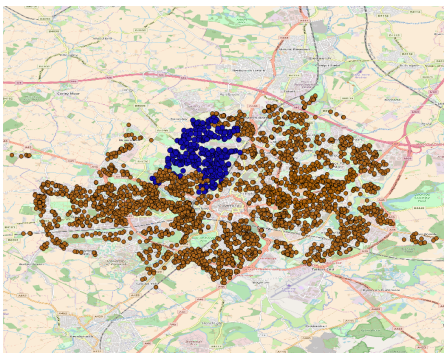
(a) Checkerboard holdout: Training points (black grid) and testing points (white grid).



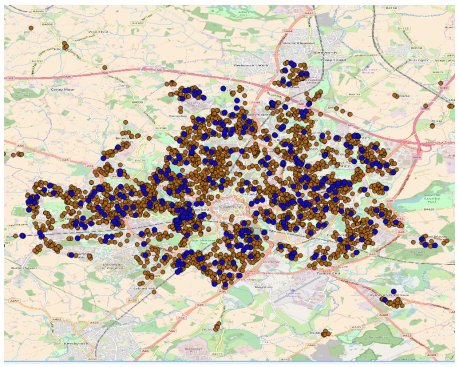
(b) Spatial k-fold with dead-zone radii: Yellow points are removed from the training set.



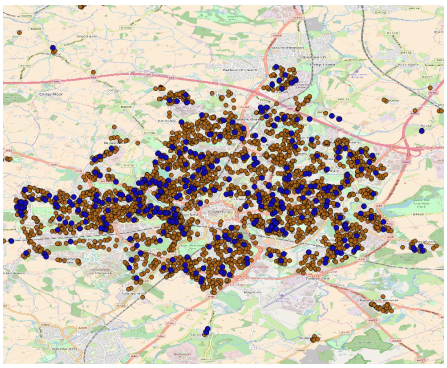
(c) Blue points are the training set and brown points are the test set, spatial K=1.



(d) Blue points are the training set and brown points are the test set, spatial K=2.



(e) Blue points are the training set and brown points are the test set, standard K=1.



(f) Blue points are the training set and brown points are the test set, standard K=2.

Figure 5.4: A comparison of all sampling techniques.

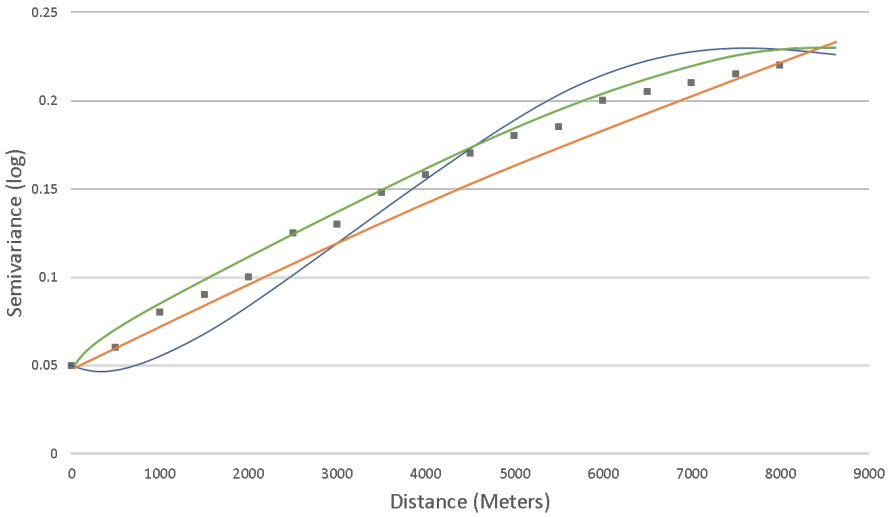


Figure 5.5: A graph of the three best kernels for a road distance matrix.

Table 5.2: Selected hyperparmeters for all experiments (1)-(6) with dead-zone 10 fold cross validation.

	Euc. Distance (Exp. 1)	Mink. Distance (Exp. 2)	PD Road [136] (Exp. 3)	Road Distance (Exp. 4)	Travel Time (Exp. 5)	Comb. Matrices (Exp. 6)
Nugget	0.03	0.003	0.0035	0.018	0.0015	0.008
Sill	0.07	0.03	0.02	0.03	0.05	0.05
Range	20000	20000	15000	15000	30	30000
Kernel	Mat	Mat	Mat	Gaus	Sph	Sph

5.5.4 Variogram construction and Ordinary Kriging

Let $s \in \mathbb{R}^d$ be a single location representing a house in a d-dimensional Euclidean space and, suppose that the house price $\mathbf{Z}(s)$ at spatial location s is a random quantity. Then, let s vary over index set D, which is a subset of $\mathbb{R}^d$ ($D \subset \mathbb{R}^d$), so as to generate the random process $\mathbf{Z}(s) : s \in D$.

For each experiment (6 in total) and each sampling technique (4 in total), a new variogram is produced together with a parametric model (kernel); see 'Variogram Construction' in Figure 5.3. The maximum distance and lag classes are empirically selected. The nugget, sill and range are selected by ordinary least squares (OLS). For each fold in a k-fold sampling technique, a new variogram is estimated. By means of an example, Figure 5.5 graphically displays the vari-

ogram for the first fold of experiment (4) (restricted road distance metric) with its three best performing kernels: Gaussian, spherical and Matern (in improving order). I undertake two approaches to selecting the best variogram: (1) The user empirically selects the kernel; (2) A maximum likelihood estimator (MLE) selects the best kernel [77]. I find that the empirical fitting approach, although lengthy to undertake, produces in all cases a matching or better predictor result. Hence, Section 5.5.6 reports the optimal results with empirical fitting for all sampling techniques as well as MLE for KCV as evidence that I selected the best approach. Table 5.2 provides the selected parameters and hyperparameters for each experiment with my most realistic sampling approach - S-KCV. It can be seen that the kernel used can change between each experiment, this is because I select the kernel which produces the best Kriging result for each experiment. The kernels show that different distance matrices can make a significant difference to the parameters and weightings of an optimal Kriging predictor. Given that I provide the best result, irrelevant of the kernel, I am providing a more robust like-for-like comparison than I would if I just selected one kernel for all experiments. I believe that this avoids overly optimistic results for one or two experiments and pessimistic results for the remainder.

5.5.5 Validation

Three validation metrics are utilised, (1) r^2, (2) Root Mean Squared Error (RMSE) and (3) Mean Absolute Percentage Error (MAPE) (see Section 2.5.1 for their definitions).

5.5.6 Results and analysis

A summary of all results are recorded in Table 5.3, which provides the validation results for each experiment (1-6) for all validation techniques (k-fold, Checkerboard, SS-KCV and S-KCV). All values in bold represent the experiment which provides the best house price predictor for each sampling technique. If more than one experiment is selected for one sampling technique, then all re-

sults between are statistically insignificant based on a t-value of 0.05 on a paired t-test, and hence all are optimal. It can be seen that the prior state-of-the-art (Euclidean and Minkowski) consistently under-perform compared with the urban road distance and travel time based models. For example, Euclidean based Kriging delivers an r^2 of 0.23 compared with a combination of road distance and travel time of r^2 of 0.56 (>x2 goodness of fit) on the most pessimistic/realistic sampling technique (S-KCV). In addition, I note that by considering the shortest path with restrictions (i.e., experiments (4)-(6)), unlike the current state-of-the-art in isomap (experiment (3)), I am able to find a statistically improved house price regression in 3 out of 4 sampling techniques.

Notably, the significance of the improvements between my new approaches (experiments (4)-(6)) compared to Euclidean distances increase as the sampling technique becomes more pessimistic. This is intuitive because in S-KCV, a Euclidean dead-zone is utilised to penalise the over bias caused by SAC. Additionally, my novel approaches take account of a more sophisticated SAC which better infers the covariates of an urban environment and hence is less affected by the assumption of independent and identically distributed (i.i.d) random variables in KCV.

As previously discussed (Section 5.5.4), Table 5.4 presents the results for all experiments with a maximum likelihood estimator. These are inferior to the empirical approach, hence I opted to undertake all experiments with the empirical approach; these results are shown in Table 5.3. Table 5.5 emphasises this point by reporting that my empirically selected kernels produce improved urban house price Kriging predictors compared with the MLE approach undertaken in [35].

Table 5.3: Results from four validation techniques: 10-fold cross validation, spatially stratified 10-fold cross validation, checkerboard holdout and spatial dead-zone 10-fold cross validation.

Results Tables							
	Previously Implemented Techniques for Comparison			**Newly Defined Techniques**			**Optimistic**
	P=2 (Euclidean) (Exp. 1)	P=1.6 (Minkowski) (Exp. 2)	No Road Restriction (Exp. 3)	Road Distance (Exp. 4)	Travel Time (Exp. 5)	Combined Matrices (Exp. 6)	
	10-Fold Validation						
r^2	0.81±0.3	0.8±0.18	0.79±0.03	**0.82±0.06**	0.81±0.06	**0.82±0.04**	
RMSE	55177±13034	74786±29266	65088±15481	**57322±18958**	59294±12830	**78158±21742**	
MAPE	17.9±1.1%	24.5±6.9%	20.7±1.74%	**21.5±1.6%**	18.1±1.9%	**25.2±1.7%**	
	Spatial 10-Fold Stratified Validation						
r^2	0.42±0.21	0.44±0.26	**0.47±0.34**	**0.46±0.24**	**0.46±0.17**	0.44±0.25	
RMSE	87081.2±68889	87539±78597	**78744±36831**	**71601±62217**	**75905±68296**	77839±68127	
MAPE	32.3±21.2%	30.4±22.5%	**26.8±11.5%**	**25.7±10.02**	**26.5±12.3**	26.6±13.6%	
	Checkerboard Stratified Validation						
r^2	0.44	0.46	0.5	**0.51**	**0.51**	**0.52**	
RMSE	82972	81940	76850	**72770**	**75226**	**74816**	
MAPE	26.7%	26.2%	24.9%	**23.8%**	**26.1%**	**25.3%**	
	Dead-Zone 10-fold Cross Validation 20 Meters						
r^2	0.23±0.13	0.29±0.32	0.53±0.16	0.5±0.09	0.4±0.15	**0.56±0.05**	
RMSE	97079±18491	100201±39526	85770±11052	87730±21736	97892±22792	**85413±9138**	**Pessimistic**
MAPE	31.2±3.4%	34.7±4.6%	28.3±2.5	26.5±3.02	31.3±3.9	**27.2±2.9**	

Table 5.4: Maximum likelihood results with dead-zone spatial k-fold cross validation.

	Euclidean Distance (Exp. 1)	Minkowski Distance (Exp. 2)	PD Road [136] (Exp. 3)	Road Distance (Exp. 4)	Travel Time (Exp. 5)	Combined Matrices (Exp. 6)
r^2	0.187	0.236	0.431	0.413	0.327	0.457
RMSE	102155.62	108238	91047	94655.37	104157	92051
MAPE	32.60	33.25	36.01	29.62	27.79	28.04

Table 5.5: A comparison of the results from [35] (Contribution 1) with those from this contribution using 10-fold cross validation.

	P=2 New	P=2 [35]	P=1.6 New	P=1.6 [35]
r^2	0.801	0.663	0.8	0.6901
RMSE	55177	58913	74786	57013
MAPE	17.9%	18.12%	24.5	17.895%

Overall one can see that my isomap approach can, in some cases, deliver a goodness of fit which is twice as good as results from an approach using Euclidean distance. This statistically significant outcome highlights the potential of using restricted road distance, travel time and non-Euclidean distance matrices, in urban studies and in other geostatistical applications such as restricted stream distances.

Isomap is representative of a network's global structure, and is theoretically understood across disciplines. Local isometric embedding on the other hand, attempts to preserve the local geometry of data, these methods include sparse matrix computations that speed up calculation and utilize local geometry and Euclidean distances in a network, which may otherwise be non-Euclidean globally. Given that I have utilized the commonly understood global approach, further research would include testing against local isometric embedding, especially if one were interested in producing real-time applications which require a low computational complexity.

5.6 Final Remarks

Through the use of a practical urban modelling case study, I demonstrate that variogram functions do not always remain valid with non-Euclidean distance inputs, and therefore establishing the validity of each distance function becomes essential. Using isomapping - a method for nonlinear dimensionality reduction - I show that it is possible to produce PD Euclidean distance metrics, and as a result valid variogram functions.

In contrast to previous research, I demonstrate that shortest path link-based road distances do not always improve the output of geostatistical models compared with Euclidean-based approaches. However, road networks which consider real-world restrictions, such as one-way systems, congestion and the presence of traffic lights can significantly improve modelling accuracy. Two such approaches presented in this research are travel time and a combination of restricted road distance and travel time, both of which account for a greater number of factors than road distance alone.

More specifically, a newly defined isomap approach is presented, which shows that road distance and travel time can both be more accurately modelled against a PD approximation of both, compared to Euclidean, Minkowski and link-based approaches [35, 136]. In some cases this provides a goodness of fit value which is twice as good as state-of-the-art approaches.

Furthermore, an extensive comparison of spatial CV techniques is conducted, in which I conclude that KCV does not accurately estimate how well a model generalises to unseen data in a spatial setting: S-KCV is shown to be a more appropriate sampling technique for CV.

I highlight that using an inappropriate validation sampling technique can lead to an incorrect selection of prediction models. In the case study that I present, the results for my combined road distance and travel time method is significantly better with SAC removal than with standard KCV. The results show that restricted road distance and travel time predictions produce a sta-

tistically improved house price predictor with an r^2=0.56; this compares with a Euclidean-based approach which achieves a result of r^2=0.23 in the case of sampling with a pessimistic/realistic dead-zone KCV technique (S-KCV).

Further avenues of research include the introduction of covariates for an optimal AVM, the production of a restricted road distance and travel time kernel for urban variogram modelling and an improved estimate of a combined road distance and travel time metric.

In the following chapters, I will firstly introduce a state-of-the-art urban spatial CV method for estimating the generalisation performance of spatial predictive models along the range of interpolation to extrapolation scenario's. Thereafter, Chapter 7 puts forth a set of answers to the research questions (RQ) posed in Section 1.1, matches those answers with my results, and discusses the implications of my work to urban science, geostatistics and real estate. Finally, Chapter 8 concludes all of my findings and puts forth a set of research avenues that are opened up by this thesis.

CHAPTER 6

Road Distance and Travel Time Cross-Validation for Urban Models

Chapters 4 and 5 confirm that physical and social processes in urban systems are inherently spatial and hence data describing them contain spatial autocorrelation (SAC) that need to be accounted for when modelling. Similarly, standard k-fold cross validation (KCV) techniques that attempt to measure the generalisation performance of machine learning and statistical algorithms also need to take account for such spatial dependencies. For example, if one were not to take account of spatial dependencies between training and test sets, then an overestimation of generalisation performance to unseen data may occur.

The current literature introduces a number of methods to take account for such dependences between the training and test sets, examples include: blocking [111] cross validation and spatial k-fold cross validation (S-KCV) [106]. However, the physical barriers and complex network structures which make up a city's landscape means that even these methods can be inappropriate. This is again due to the assumption that mobility, and hence SAC is Euclidean in nature, which is not appropriate in cities. To overcome this problem, I propose a new road distance and travel time k-fold cross validation method and I show how it outperforms the prior art at providing better estimations of generalisation performance to unseen data. This chapter addresses **RQ3** fully, and the results are taken from Publication **6**.

6.1 Introduction

As previously dicussed, cities are growing; it is expected that 5 billion people will live in urban spaces by 2030 [12, 98, 114]. Consequently, models for sustainable and high quality urban life are being built, notably spatial models e.g., [136] and [37]. Given the increased reliance on these, sometimes 'blackbox' models [17], it is essential that we understand how well they perform for predictive and explanatory purposes. As such, it is necessary to correctly estimate how well a model generalises to unseen locations, especially with applications where data are spatially inconsistent or sparse. Cross validation is the typical technique utilised to report such estimates - it is essential that these methods take account for internal dependencies, most notably - space.

Specialist cross validation techniques have been put forth to estimate such dependency structures for example, S-KCV [106], blocking [111] and stratified sampling [37]. Each of these methods attempt to account for the SAC between test and training points, which traditional cross validation (CV) methods do not.

Specifically, S-KCV attempts to remove SAC by implementing a Euclidean 'dead-zone' area around all test points, such that all training points that lay in these areas are removed - this method was utilised for CV in Chapter 5. However, I argue that Euclidean distances may not be appropriate for urban systems, Chapters 4 and 5 provide such intuition.

6.1.1 Contributions

In this chapter, I introduce a new spatial k-fold cross validation method, termed road distance and travel time k-fold cross validation (RT-KCV), which constructs and utilises *road network and travel time* dead-zones. The key contributions and benefits of RT-KCV are: (1) state-of-the-art estimates of the generalisation performance of any spatial urban model across the interpolation-extrapolation range of application scenarios; (2) significant improvements in

efficiency of dead-zone training point removal when compared with the current state-of-the-art (S-KCV [106]) and (3) improved performance in capturing and removing urban SAC. I demonstrate these contributions across two large-scale urban datasets and three different scenarios of interpolation-extrapolation. I also provide an extensive experimental comparison across multiple CV techniques and offer a systematic way to choose the dead-zone distance.

6.1.2 Chapter structure

Section 6.2 provides a full description of why and under what settings SAC removal is required. Section 6.3 reviews related approaches for SAC detection and generalisation performance (i.e., a model's ability to generalise to an unseen location). Thereafter, Section 6.4 redefines spatial cross validation for urban spaces, utilising a unique set of restricted road, travel-time and combined distance dead-zones. Section 6.5 then introduces two urban datasets over three (interpolation - extrapolation) modelling settings with the purpose of comparing the estimated generalisation performance of several validation methods - KCV, S-KCV, R-KCV, T-KCV, RT-KCV and blocking KCV. Finally, Section 6.6 offers some final remarks about the findings in this chapter.

6.2 Problem Definition

Cross-validation splits a dataset into two subsets - a **training set** with which a model is established and a **validation test set** against which the resulting model is evaluated [117]. The main purpose of cross validation is to detect overfitting and estimate how well a model will generalise to unseen data - sometimes referred to as a **ground truth test set**. Specifically, KCV repeats the process k times, validating on all the disjoint subsets of the dataset. Since urban problems are inherently spatial in nature, a chosen cross-validation method should be able to accommodate and mimic different spatial scenarios such as interpolation, extrapolation or some combination of the two.

For example, if the aspiration is to interpolate (estimate an unknown value from within a known domain), then traditional cross validation is satisfactory. The reason for this is that all unknown values in the ground truth test set will contain the same (or similar) spatial autocorrelation with the training set as the points that have been held-out by cross-validation in the validation test set. If this is the case, the cross-validation estimate of how well the model will perform (model generalisation) will be accurate.

However, if the purpose is extrapolation (to estimate an unknown value outside of a known domain), then the cross validation method must produce a validation test set which contains less or no SAC with the training set, in order to simulate the unknown out-of-range value. In all settings, other than pure interpolation, traditional CV, which assumes independent and identically distributed (i.i.d) random variables, is over-optimistic i.e., overestimates the generalisation performance of the model.

In this thesis, SAC_{train} and SAC_{test} refer to the SAC within the training and validation test sets respectively. $SAC_{traintest}$ defines the SAC between the training and validation test sets [26]. Removing $SAC_{traintest}$ to improve the estimate of a model's generalisation performance requires an understanding of how dependencies are structured and unfold in geographic space. Typically it is assumed that spatial dependence is Euclidean in nature, but in most urban settings natural or man-made restrictions (e.g., one-way road systems) violate this assumption. As such, I hypothesise that road distance, travel time and a combination of both are better able to infer urban SAC.

6.3 Related Literature

6.3.1 Spatial autocorrelation (SAC)

SAC describes the correlation of all observed variables to each other in a spatial dataset. This correlation can be explained solely by geographical proximity [66]. SAC was first influenced by the central place theory [21], which in itself was

inspired by theories of proximity and nearness [123]. Later, SAC and Moran's I were developed by [25, 96], and gave rise to a series of measures, such as Getis and Ord's G_i statistic [57] and Matheron's $1/\gamma$ (inverse of the semivariogram) [88].

Commonly, SAC is measured to (i) test model mis-specifications [24], (ii) measure the strength of spatial effects on a variable, (iii) test for spatial stationarity, heterogeneity or clustering, (iv) detect distance decay and (v) identify outliers and design spatial samples [2, 50, 55]. In this work, I remove SAC between training and validation test sets in order to better estimate the generalisation of my models in different settings. With the exception of research by [68, 134], little research has considered and utilised the different sources of SAC (SAC_{train}, SAC_{test} and $SAC_{traintest}$). No prior works have considered SAC using non-Euclidean distances for the purpose of estimating a model's ability to generalise to unseen data.

6.3.2 Model generalisation

The primary methods utilised to estimate the generalisation performance of a model to unseen data are holdout cross validation and k-fold cross validation.

Hold out

Holdout cross validation simply partitions input data into two (mutually exclusive) subsets; training and test/holdout. Typically, holdout cross validation assumes the input data to be i.i.d random variables, which is inappropriate in applications of data containing spatial, temporal, grouping and hierarchical autocorrelation [111]. As such, 'blocking' holds out autocorrelated strata's, one such example is checkerboard holdout, which splits the input dataset based on a user-defined spatial chess-board [37] to reduce SAC. Blocking holdout cross validation (1) only trains on a proportion of the available data, (2) is agnostic to the specific task at hand (interpolation vs extrapolation) and (3) contains SAC at each strata border.

K-fold Cross Validation

K-fold cross validation partitions data into k subsets, performs analysis on k-1 (training) subsets, and validates the analysis on the remainder. The process is repeated k times, where the test set is different each time. The validation results between each fold is averaged to reduce outlier bias [72]. The most typical cases of k-fold cross validation are k=10 and k=n (Leave One Out), where the latter model trains on the largest set of data possible, but is time-consuming on large datasets [46]. Traditional KCV withholds the central independence assumption which, as discussed, can provide optimistic estimates of generalisation performance [76, 111].

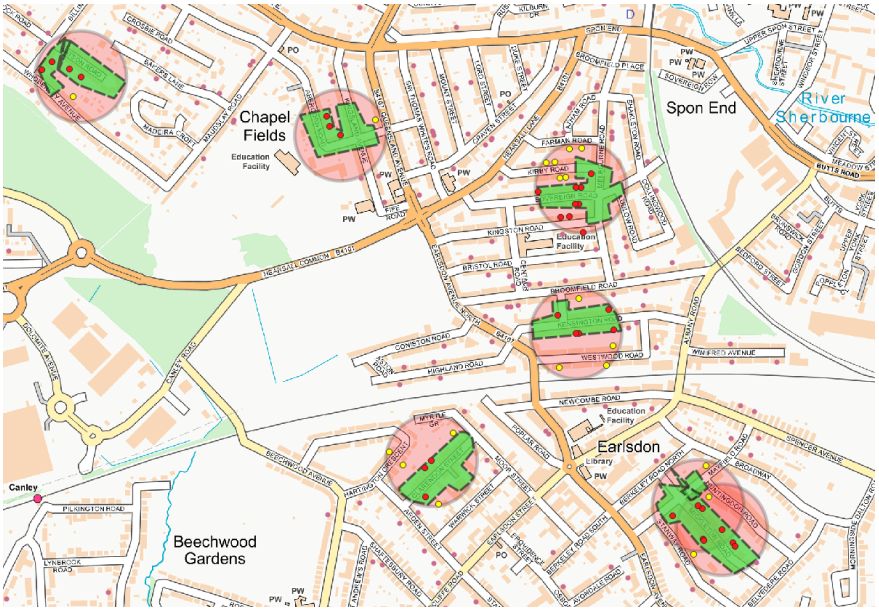
As such, Geostatisticians are critical of cross validation for confirmatory data analysis with dependent data [32]. Spatially aware cross validation methods have hence been proposed to break the dependence between the training and testing set. The most notable of these methods is S-KCV, which estimates a predictor's performance by first implementing traditional k-fold cross validation and second, removing all training points within an empirically designed Euclidean dead-zone from all test points [106]. Additionally, [78] proposes a special case of S-KCV termed spatial leave one out (SLOO), which computes a threshold distance equal to the range of residual spatial autocorrelation in order to promote spatial independence between all points. As a method for estimating how well a model will generalise to unseen data, key drawbacks of this approach are (1) the removal of valuable training points in each dead-zone, (2) the lack of an established (or even an ad-hoc) approach to choosing the dead-zone radii, (3) the disregard toward the specific nature of the task in hand (interpolation to extrapolation) and (4) the assumption that a Euclidean distance is the most appropriate function for dead-zones. As I will show, my RT-KCV method addresses all of these drawbacks.

Finally, blocking k-fold cross validation is an alternative, non-random sampling technique for validation, where the held out data lays inside some spatially defined strata [95]. The benefits of this approach over other k-fold cross

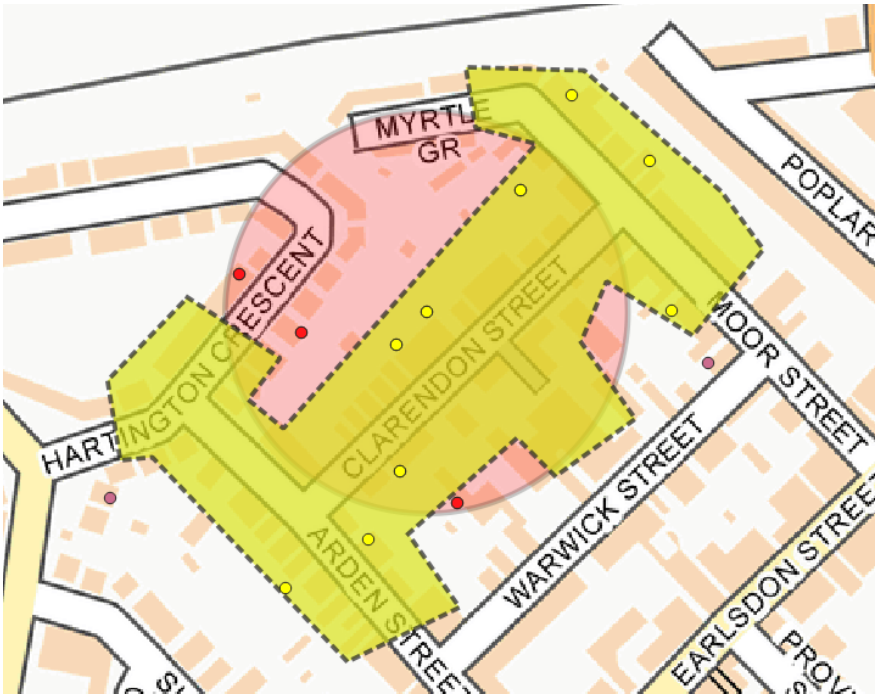
validation techniques is it's ability to simulate unseen values in unseen areas. However, a strong understanding of the spatial processes in a dataset are required. For example, some datasets contain a mixture of dense and sparse areas which can result in overfitting to one geographic area. Approaches to overcome this problem include equal frequency spatial strata's [35] or irregularly arranged regular or irregular blocks [111]. Further challenges with this method include: (1) the time-consuming and ad-hoc nature of setting up cross validation for new datasets; (2) the poor fit to problems involving interpolation and extrapolation; (3) the SAC present at block borders and (4) ad-hoc choices for the shape, size, placement and regularity of the blocks. As I demonstrate, my proposed RT-KCV method overcomes all of these issues.

6.4 Road and Travel Time Validation

RT-KCV is a spatial dead-zone technique which, in a similar way to S-KCV, constructs an area around each test point from which all training points are removed. Unlike S-KCV, RT-KCV produces contiguous, non-convex dead-zones from a combination of restricted road distance and travel time matrices. The purpose of this method is to better capture spatial autocorrelation in access-restricted areas such as cities. Figure 6.4(a) visualises these differences where road (in green) and Euclidean (in red) dead-zones are compared. The main idea behind RT-KCV is that road distance and travel time dead-zones contain more SAC than Euclidean ones. Hence, RT-KCV dead-zones are more efficient by design - that is, more SAC can be removed while removing fewer training points. Figure 6.4(b) illustrates this with a real example where the road distance (yellow) dead-zone is larger than the Euclidean (red) dead-zone, but removes fewer (and different) points. The points that have been removed by road distance are physically more accessible to the test point being considered, which for many urban applications implies higher SAC removal. I show that this is the case in two real world urban datasets and conjecture that this generalises to a plethora



(a) A subset of test points with Euclidean (in red) and road distance (in green) dead-zones.



(b) A single test point with a road distance (in yellow) and Euclidean (in Red) dead-zone, showing that road distance points based on accessibility.

Figure 6.1: An example of road distance versus Euclidean dead-zones.

of urban applications driven by human behaviour such as evaluating impact of green space, designing algorithms for car sharing, predicting house prices and designing methods to improve traffic flow. The definition below and Algorithm 4 describes the entire process of the combined road distance and travel time RT-KCV method, which is complemented by comparison with the existing state-of-the-art (S-KCV, blocking and KCV) and additional variants; R-KCV that only considers road distance, and T-KCV that only considers travel time. Figure 6.2 provides a flow diagram of the entire experimental validation process for all spatial k-fold methods — S-KCV, R-KCV, T-KCV and RT-KCV. The resulting model is Kriging-based and the validation metric is Normalised Root Mean Squared Error (NRMSE).

Algorithm 4 The RT-KCV algorithm.

Require: ν, S, $\mathcal{A}$, ρ, $\mathcal{M}$
1: $RT \leftarrow \alpha_1 * R + \alpha_2 * T$ ▷ Create RT Matrix
2: **for** RT **do** ▷ Set up CV with RT distance
3: **for** i $\leftarrow$ 1 to K **do**
4: $\mathcal{H} \leftarrow \cup_{s_k \in \nu_i}\{s_j \in S | e(c_j, c_k) \leq \rho\}$ ▷ Remove data
5: $\mathcal{F} \leftarrow \mathcal{A}(S \backslash \mathcal{H})$ ▷ Build model
6: **for** $s_k \in \nu_i$ **do**
7: $\hat{y}[k] \leftarrow \mathcal{F}(x_k, c_k)$ ▷ Prediction
8: **end for**
9: **end for**
10: **return** $\hat{y}$, NRMSE ▷ Validation results
11: **end for**

Definition. Assume a data point $s_i = (x_i, y_i, \mathbf{c}_i)$ where $x_i \in \mathrm{R}^D$ is a feature vector, $y_i \in \mathrm{R}$ is the response/target and $\mathbf{c}_i \in \mathrm{R}^2$ is the geographical coordinate vector of the ith data point in dataset $\mathbf{S}$=$\{s_1, s_2, \ldots, s_n\}$. Additionally, consider a set of distance matrices $\mathcal{M}$ = {Road Distance (RD), Travel Time (TT)}. I define $\rho \in \mathrm{R}^+$ to be the dead-zone radius and $\nu = \{\nu_1, \ldots, \nu_k\}$ to be the set of KCV folds. Vector $\hat{\mathbf{y}} \in \mathrm{R}^n$is the predicted response values from model $\mathcal{F}$ (in my case - Kriging). Additionally, $\alpha_{1,2}$ are some user defined weightings to calculate to what extent road distance and travel time form a basis of my RT matrix. Finally, a validation metric (in this case 'NRMSE') is selected.

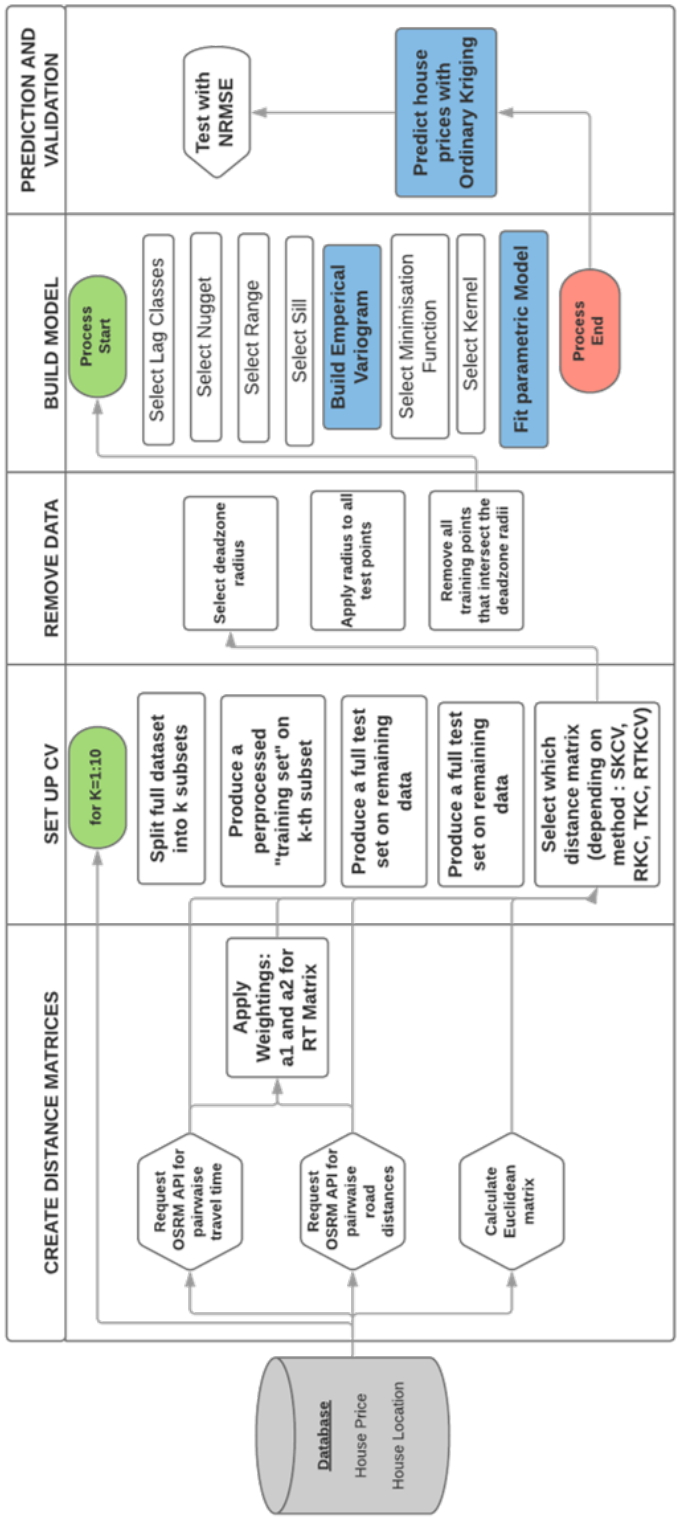


Figure 6.2: A flow diagram of S-KCV, R-KCV, T-KCV and RT-KCV algorithm.

Determining the Optimal Dead-Zone. RT-KCV and variants have a free parameter - the dead-zone distance. The dead-zone radius is typically user defined. However this directly defines the amount of SAC that will be removed and hence implicitly defines the setting (interpolation-extrapolation) a model is expected to be in and hence how accurate the estimated generalisation performance will be. In this research I propose a dead-zone heuristic in order to provide a single dead-zone distance for any KCV method which will approach the ground truth value. The heuristic calculates the average pairwise distances between all points in the training and ground-truth test sets (termed the *similarity matrix*). The second step of the heuristic finds the 'maximum separation distance' (d_{max}) taken from the training set's semivariogram to provide an upper bound of distances. All training/test points which have a distance greater than d_{max} are removed from the train/test distance matrix to produce a new 'SAC only' distance matrix (μ_{tt}). This provides a set of train and test points which are assumed to be correlated. Thereafter, one would select the dead-zone distance based on the following heuristic that was found to perform well across settings and datasets

$$Distance = \begin{cases} 0, & \text{if } \frac{\mu_{tt}}{\mu_{tr}} \leq 1. \\ \mu_{tt}, & \text{if } \frac{\mu_{tt}}{\mu_{tr}} > 1 \text{ and } \mu_{tt} < d_{max} \\ d_{max}, & \text{otherwise.} \end{cases} \tag{6.1}$$

where μ_{tt}, μ_{tr}, μ_{te} are the average distances in the validation train/test, train and validation test sets respectively. Once the distance is selected, one would find how many points are removed and then use this value to determine the dead-zone area for any method (R-KCV, T-KCV, RT-KCV, S-KCV). The output of this heuristic I term the *'mean operating point'*.

6.5 Urban Case Studies

Given that (non-Euclidean) distance is shown to be the single most influential variable in urban house price predictions [35, 37], my first case study builds a valuation model with no covariates on a set of 3,413 residential sold house prices in Coventry, United Kingdom (UK) projected to 2017. My second case study utilises historic traffic flow information on 711 sensor locations in Birmingham, UK.

6.5.1 The base Kriging predictor

For both case studies, I consider Ordinary Kriging — a spatial predictor which accounts for spatial covariance based on observed pairwise distances. I use Ordinary Kriging as a simple and widely utilised spatial statistical model in order to demonstrate the benefits of RT-KCV. The method of interpolation with Ordinary Kriging is defined in Section 2.3.2.

Defining Non-Euclidean Dead-Zones

The Open Street Routing Machine (OSRM) provides the distance and time it takes to travel from one location to another by car through a simple to use API. Their link-based algorithm utilises a set of restrictions defined in OpenStreetMaps (OSM) (see Table 3.1.1). From their API, one is able to calculate an nxn distance matrix for all points. My combined RT-KCV approach is calculated such that travel time and road distance are both normalised between 0 and 1 and then summed with a weighting (0.5 for both case studies). This weighting is empirically selected given that both road distance and travel time perform better at different stages of the variogram. More details on this method are discussed in Section 3.1.1. I speculate that a future avenue of research is to build a heuristic/metric which can optimise these weightings in In Section 6.6.

6.5.2 Validation

I describe the comparison and evaluation of the proposed KCV methods against the state-of-the-art. The primary purpose of any (cross-)validation procedure is to estimate as best as possible a model's generalisation performance to unseen data. As such I propose three settings (interpolation, extrapolation and inbetween) over two case studies (house price and traffic flow prediction), each with a number of KCV methods (KCV, S-KCV, R-KCV, T-KCV, RT-KCV and blocking KCV). In order to evaluate the validation techniques, I compare each of these against a 'ground truth' value - performance on the unseen data (the ground truth test set). This is repeated across a set of six simulated real-world scenarios that are visualised in Figures 6.4(a)-6.4(f). The cross validation method which performs closest to my ground truth is the best performing. For robustness, I keep the same ground truth test in all experiments within a case study and the same validation test sets for all cross validation approaches.

I also compare my method against the most popular competitor approach - blocking cross validation [111]. The blocking approach uses 10 folds and is set up such that, for each fold, 10 random points within the training area are selected and a square block grows out so that all blocks have the same number of points in them ($\pm$ 1). All of the blocks in a fold then sum up to the same validation test set size. This provides a fair comparison for all cross validation methods. The test sets are 256 and 72 for house prices and traffic flow respectively. See Figure 6.3 for a visualisation of blocking on a subset of 3 folds.

In order to account for any variability due to the choice of the ground truth test set I re-sample multiple ground truth test sets for my non-interpolation settings (settings B and C; defined in Sections 6.5.3 and 6.5.4 for each case study). This does not require re-running any of the validation procedures as it only provides us with the stability of the ground truth test performance. Below I present three approaches to validate my KCV methods against the ground truth.

Model Validation: the Kriging model is validated against the NRMSE which

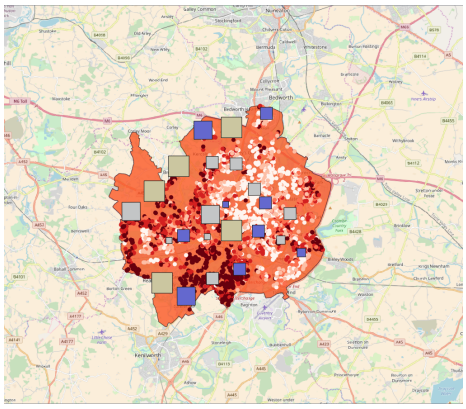


Figure 6.3: Blocking KCV with equal test sets.

intuitively takes the square root of the sum of the mean squared errors and is then normalised by the difference of the y values:

$$NRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{y_{max} - y_{min}}. \tag{6.2}$$

Convergence to the Ground Truth: this method tests how many points must be removed from a training set to achieve 50, 80 and 100% of the ground truth's NRMSE from cross validation. The purpose of this is to find out which method can obtain a 'true' NRMSE with the fewest training points removed by dead-zones. I state that the method with the largest training set at the ground truth threshold is the most effective.

Distance from Ground Truth to Estimated Dead-Zone: Section 6.4 describes a method to determine the dead-zone area. This validation measure simply calculates the difference between the NRMSE at the optimal dead-zone ('mean operating point') with the ground truth. The KCV method which has the smallest distance is deemed the most effective.

6.5.3 Case study 1 - automated valuation model

The house price data is described in Chapter 3 and considers 3,669 properties in Coventry. I consider 3 settings - pure extrapolation, mixed interpolation/extrapolation and pure interpolation - to test my newly defined methods across a range of experiments (see Figures 6.4 (a)-(c)).

Setting A - Pure Extrapolation: I train on all data that sit within the Office of National Statistic's (ONS) classified Built Up Area (BUA), accounting for 3,413 houses. The remainder are removed for testing in my ground truth test sets which account for 256 points. To simulate extrapolation fully, I confirm that my train and hold out sets are not correlated (i.e. $SAC_{traintest} \sim 0$). Again, a standard Moran's I test is conducted between both datasets showing a weak spatial relation such that $I_{observed} = 0.020206$ and $I_{expected} = 0.019014$. As such, I confirm that my method can be tested against the split data for extrapolation generalisability. All KCV approaches utilise the same test spaces which are also the same size as the ground truth and blocking KCV method.

Setting B - A Mixture of Interpolation and Extrapolation: I train on data that sit within the Coventry BUA only. For my ground truth scenario, half the test set sits within the BUA and half sits outside, thus the training set consists of 3,291 houses.

Setting C - Interpolation: I train on data that sit within the Coventry BUA. For my ground truth scenario, all the test points lay within the BUA, accounting for a training set of 3,163 houses.

Results

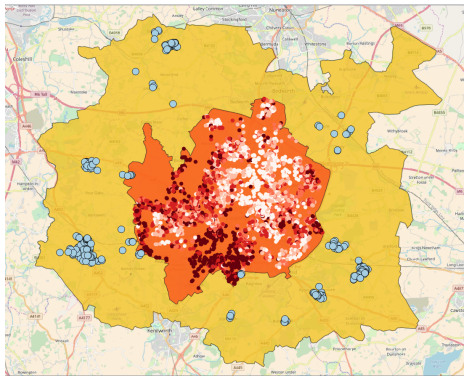
All methods in all settings have a test set of 256 points for comparison. In addition, each CV method contain the same test points for each setting. Figures 6.5(a)-(c) show the NRMSE value for each cross validation method (KCV,

S-KCV, R-KCV, T-KCV and RT-KCV). In addition, each graph shows an equal training set random removal KCV approach, blocking KCV and a ground truth NRMSE. Each KCV method is run over 10 folds and repeated 10 times (100 folds in total), showing that RT-KCV consistently outperforms all other approaches in all settings (that is it approaches the ground truth with fewer points removed). Notably, RT-KCV requires only 8 points to be removed to ensure the same SAC removal as 201 points for S-KCV in my interpolation setting. In addition, Table 6.1 shows that RT-KCV consistently generalises 50%, 80% and 100% of the ground truth with fewer points removed than any other method. Finally, my dead-zone radius heuristic estimates that 3,170, 2,003 and 0 points need to be removed to obtain an estimate of generalisation performance for extrapolation, mixed and interpolation respectively. Once implemented, I determine the difference in the estimated NRMSE values (0.11162 ,0.128 and 0.104) and the ground truth values (0.1125, 0.1225974 and 0.1135) which are relatively small compared to S-KCV and blocking. A t-test shows that for all settings, the number of points that are removed from the training set are significantly less with my new RT-KCV approach compared with the previous state-of-the-art, with a t-value of 0.01.

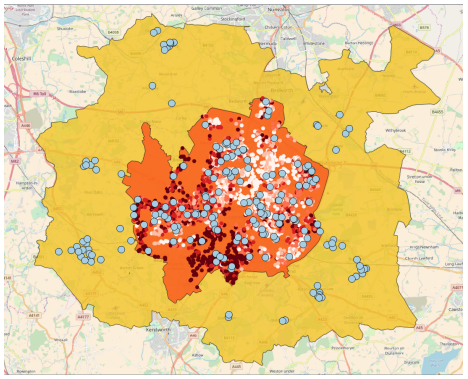
6.5.4 Case study 2 - traffic flow prediction

My predictor considers the total average daily traffic flow between $01-01-2016$ and $01-06-2017$' for Birmingham, UK accounting for 711 sensors as described in Chapter 3.

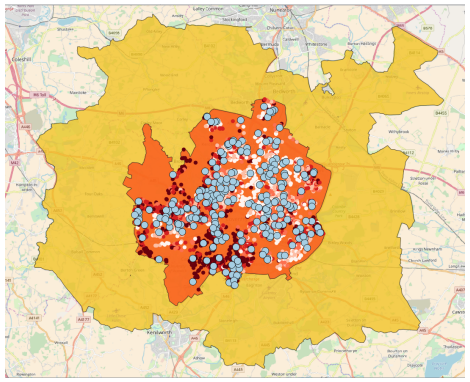
Setting A - Extrapolation: I train all data that sits within Birmingham's BUA, accounting for 711 sensors. The remainder are removed for ground truth testing. To fully simulate extrapolation, I confirm that my training and hold out sets are not correlated (i.e. $SAC_{traintest} \sim 0$). A standard Moran's I test is



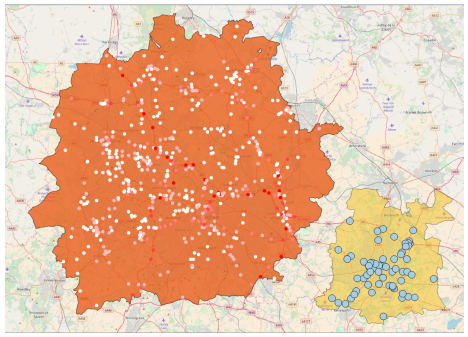
(a) A holdout method to simulate extrapolation.



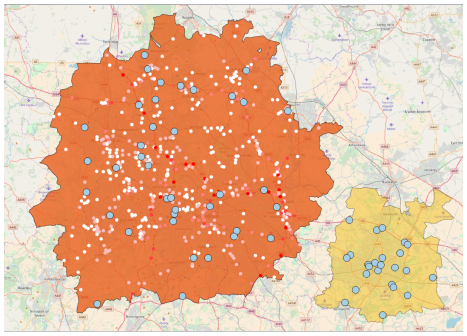
(b) A holdout method to simulate a mixture of extrapolation.



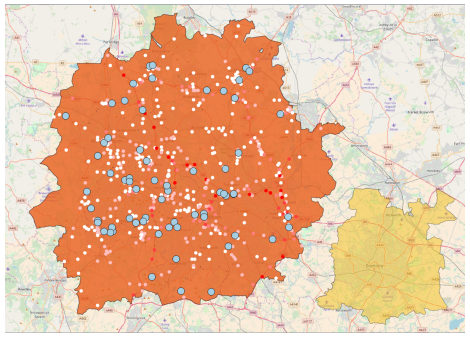
(c) A holdout method to simulate interpolation.



(d) A holdout method to simulate extrapolation.

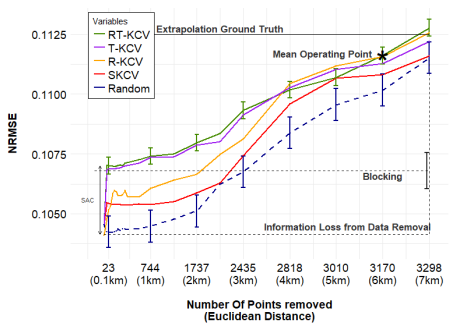


(e) A holdout method to simulate a mixture of extrapolation and interpolation.

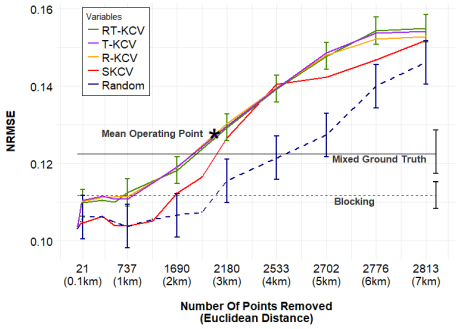


(f) A holdout method to simulate interpolation.

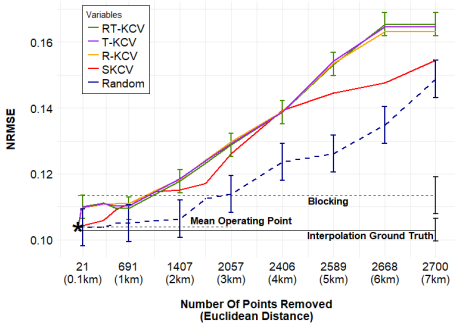
Figure 6.4: Producing a ground truth train and test set. The orange space represents the training area, the yellow space represents the ground truth test area, the blue points are ground truth testing locations and the white to red points represent the training set where the white points are the cheaper houses/lower traffic flows and red points are the more expensive houses/higher traffic flows.



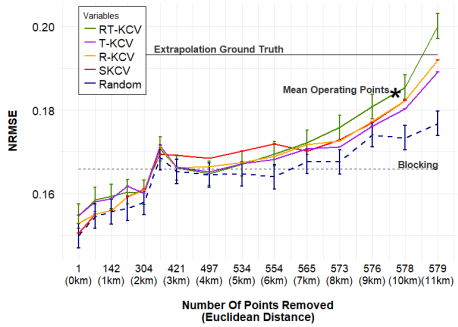
(a) NRMSE for Ordinary Kriging on Coventry House Price with all KCV Methods Compared with Extrapolation.



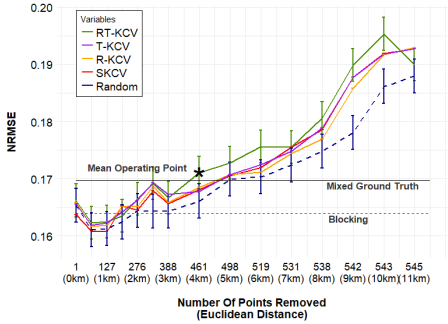
(b) NRMSE for Ordinary Kriging on Coventry House Price with all KCV Methods Compared with a Mix of Interpolation and Extrapolation.



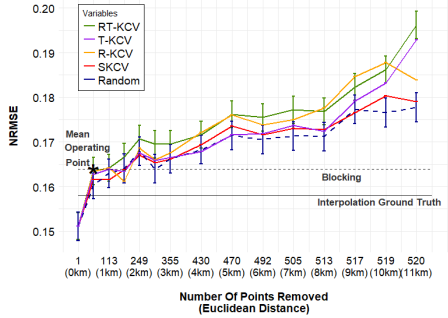
(c) NRMSE for Ordinary Kriging on Coventry House Price with all KCV Methods Compared with Interpolation.



(d) NRMSE for Ordinary Kriging on Coventry House Price with the Winning KCV versus Random KCV with Standard Error Bars.



(e) NRMSE for Ordinary Kriging on Coventry House Price with the Winning KCV versus Random KCV with Standard Error Bars.



(f) NRMSE for Ordinary Kriging on Coventry House Price with the Winning KCV versus Random KCV with Standard Error Bars.

Figure 6.5: Results graphs for both case studies: dead-zone size versus NRMSE for all KCV methods and the ground truth.

Table 6.1: Results: the number of points removed to reach a specific % of the ground truth NRMSE for each KCV technique.

Results Table										
	Real Estate Case Study					Traffic Flow Case Study				
	Random	Previous Work [106]	My Work			Random	Previous Work [106]	My Work		
	KCV	S-KCV	R-KCV	T-KCV	RT-KCV	KCV	S-KCV	R-KCV	T-KCV	RT-KCV
	Case A : Extrapolation (Train: 3412 - Test: 256)					Case A : Extrapolation (Train: 711 - Test: 72)				
100%	3298+	3298+	3254	3298+	**3274**	579+	579+	579+	579+	578
80%	3298+	3298+	**3105**	3156	**3112**	579+	**578**	**578**	**578**	**577**
50%	2850	2628	2489	2201	**2112**	576	573	573	573	**566**
	Case B : (Train: 3163 - Test: 256)					Case B : Mixed (Train: 675 - Test: 72)				
100%	2183	1931	**1420**	**1391**	**1401**	498	487	487	478	**442**
80%	2108	2006	**1270**	**1308**	**1295**	458	309	298	**276**	**276**
50%	1940	178	**9**	**8**	**8**	62	**57**	68	71	73
	Case C : Interpolation (Train: 3290 - Test: 256)					Case C : Interpolation (Train: 639 - Test: 72)				
100%	1489	201	**10**	**8**	**8**	84	72	**52**	**57**	**55**
80%	1417	199	**8**	**6**	**6**	67	60	**42**	52	**46**
50%	201	164	**4**	**4**	**4**	42	**31**	**30**	**29**	**30**

conducted between both datasets showing a weak spatial relation such that $I_{observed} = -0.008960041$ and $I_{expected} = 0.000201$. As such, I confirm that my method can be tested against the split data for extrapolation generalisability, see Figure 6.4(d) for a visual representation.

Setting B - Interpolation: I train on some of the data that sits within Birmingham's BUA, accounting for 675 sensors.

Setting C - A Mixture of Interpolation and Extrapolation: I train on some of the data that sit within Birmingham's BUA, accounting for 639 sensors.

A Competitor Case for Comparison - Blocking: My blocking approach uses 10 folds and is set up such that, for each fold, 10 random points within the training area are selected and a square block grows out so that all blocks have equal frequency ($\pm$ 1) and also sums to the same sized test set as all other experiments (72 points). I only apply this in settings B and C because setting A contains no test points within the training set.

Results

All methods in all settings have a test set of 72 points for comparison. In addition, each KCV method contains the same test points for each setting. Figures 6.5(d)-(f) show the NRMSE value for each cross validation method (KCV, S-KCV, R-KCV, T-KCV and RT-KCV). Additionally, the graphs show equal training set random removal, blocking and each settings ground truth NRMSE. Each KCV method is run 10 times and over 10 folds, showing that RT-KCV consistently outperforms all other approaches in all settings. Notably, the benefits of RT-KCV to my case study, although strong, is less significant for this case study compared with my house price case study, this can be explained by the weaker spatial correlation as seen by my Moran's I value in Section 3.3. Finally, my dead-zone radius heuristic estimates that 577, 458 and 87 points need

to be removed for extrapolation, mixed and interpolation respectively. Once implemented, I determine that the difference in the estimated NRMSE values (0.184, 0.172, and 0.1635) compared with the ground truth values (0.193265, 0.170, 0.158) are relatively small compared to S-KCV and blocking (with the exception of interpolation which is negligible). A t-test shows that for two out of three experiments (extrapolation and mixed), the number of points that are removed from the training set are significantly less with my new RT-KCV approach compared with the previous state-of-the-art, with a t-value of 0.01. In addition, Figure 6.5 empirically demonstrates a significant estimation of generalisation improvement, because one can see that the 'mean operating point' (my newly defined measure of generalisation performance) is significantly closer to the ground truth in all scenarios of extrapolation to interpolation, compared with S-KCV and blocking (the current state-of-the-art) for both case studies.

6.6 Final Remarks

The purpose of cross validation is to estimate how well a model will generalise to unseen data and unlabelled locations in spatial settings. However, standard KCV assumes all data to be i.i.d random variables and hence does not take into account the dependencies between the training and test set, which causes bias and optimistic estimates of generalisation. SAC is always present with spatial data and as such needs to be accounted for. Traditional validation approaches such as KCV omit the effect of SAC in performance estimations to unseen locations with urban datasets. To account for SAC in urban data I demonstrate that my new approach, termed RT-KCV, can be used to better estimate the generalisation ability and predictive performance of spatial models than existing state-of-the-art approaches (S-KCV). I also show that road distance and travel time can decrease the required 'dead-zone' data removal for capturing SAC in urban spaces, leading to a more efficient use of labelled datasets. Finally, I confirm that RT-KCV is a superior approach for estimating model generalisation

compared with all other CV methods.

I recommend that RT-KCV be used wherever dependence structures exist in a dataset with restricted space (such as cities), even if no structure is visible in the fitted model residuals, or if the fitted models account for such correlations (for example in Kriging). I note that standard KCV is only appropriate for pure interpolation where the internal dependence structure is present in the unknown values. Notably, I show that, for urban data, a combination of road distance and travel time capture SAC better than Euclidean distances.

Further avenues for research include: (1) developing techniques to better map SAC in other dependent datasets, such as 'stream' distances (along a river or canal); (2) optimising the operating point on the RT-KCV curve to better match the ground truth performance and (3) learning the convex combination parameters for the combined RDTT distance i.e., remove the requirement to manually select some weighting of road distance and travel time.

In the remaining two chapters, I will (1) put forth a set of answers to the research questions (RQ) posed in Section 1.1, match those answers with my results, and discusses the implications of my work to urban science, geostatistics and real estate and (2) conclude all of my findings and put forth a set of research avenues that are opened up by this thesis.

CHAPTER 7

Discussion and Applications

"The only true voyage... would be not to visit strange lands but to possess other eyes, to see the universe through the eyes of another, of a hundred others, to see the hundred universes that each of them sees"

Marcel Proust (1923), *La Prisonnière* from the *Remembrance of Things Past.*

Cities are inherently spatial; *urban proximity is related to mobility* and *restricted road networks can measure urban space*: three statements which the findings in Chapters 4-6 confirm. Additionally, non-Euclidean distances can improve (1) geostatistical urban models and (2) the estimation of the generalisation performance of a (spatial or otherwise) model for all interpolation-extrapolation scenarios.

The above summary of findings is examined in detail throughout this Chapter. Section 7.1 outlines the thesis contributions in response to the research questions put forward in Chapter 1. Thereafter, the implications of this thesis research on urban science, geostatistics and the real estate industry are considered in Sections 7.2-7.4. Finally, the potential limitations to the generalisation of this research are introduced in Section 7.5.

7.1 Answers to Research Questions (RQ)

At the start of this thesis three research questions were put forth:

1. RQ1: Which distance function best models spatial interactions in an urban setting?

2. RQ2: When, if ever, are non-Euclidean distance functions valid for urban spatial models?

3. RQ3: How should one estimate the generalisation performance of urban spatial models?

Each research questions (RQ) motivates the contributions throughout this thesis and we explore these contributions below.

7.1.1 Research undertaken in response to RQ1

RQ1 is fully answered in Chapter 4, and the results are taken from Publication 4.

Urban processes in space result in data which are not independent and identically distributed (i.i.d) random variables. Semi-variograms [33], Moran's I [96] or Getis's G [56] are just some examples of statistical measures that describe the extent of these dependencies to better allow them to be taken into account. Each of these methods have a notable commonality - distance is measured with a Euclidean function. Hence, these distance-based learning methods do not take account of the physical properties of dispersion in a city landscape, as discussed on several occasions in this thesis.

Additionally, for geostatistical interpolation (i.e., Kriging - Section 2.3.2), it is essential to ensure that existing covariance and (semi)variance functions remain valid, positive definite (PD) and conditionally negative definite (CND) respectively [39].

The work in Chapter 4 overcomes both of these issues by finding a distance function which can better measure urban space, whilst still producing a valid covariance and variogram function. This chapter shows that (1) normalised road distance, travel time and combined distances can better model urban spatial autocorrelation (SAC) in a semivariogram and (2) that the valid Minkowski distance which measures the highest similarity with non-conforming road distance, travel time and combined functions also produce the best spatial interpolation. The distance function produces an improved set of parameters and hyperparameters in the semivariogram which, in turn, causes the Ordinary

Kriging interpolation presented here to outperform the state-of-the-art (Euclidean) distance function on a real world house price case study in Coventry. As such, I have shown that there are some non-Euclidean functions which can better model urban mobility.

7.1.2 Research undertaken in response to RQ2

RQ2 is addressed in Chapter 5, and the results are taken from Publication 5.

Spatial models do not assume data to be i.i.d random variables. Minkowski pairwise distances may not be the best estimation of road distance, travel time and a combination of both. Additionally, spatial modelling has been extensively studied with a straight line, Euclidean, pairwise distance. Given these facts, there is no guarantee that any other improved non-Euclidean distance matrix (PD or otherwise) will produce a valid covariance or (semi)variance function.

Hence, Chapter 5 presents a method to approximate restricted road distance, journey time and combined matrices into an embedded lower-dimensional Euclidean space to ensure that covariance and (semi)variance functions remain valid when using urban-specific distances. For confirmation of an improved spatial interpolation, I provide a comparison of six Ordinary Kriging predictions, each with a different distance metric, employed in a real estate case study. The distance matrices utilised were neither originally Euclidean or PD, as such Chapter 5 shows, for the first time, that any non-Euclidean distance function can be mapped into a valid Euclidean function for the purpose of proximity based modelling.

7.1.3 Research undertaken in response to RQ3

RQ3 is addressed in Chapter 6, and the results are taken from Publication 6.

The main purpose of k-fold cross validation (KCV) is to detect over fitting and estimate how well a model will generalise to unseen data i.e., the expected performance of a 'ground truth' test set (defined in Chapter 6). This method assumes that the random variables in the validation test and training sets are i.i.d. However, urban problems are inherently spatial, which invalidates this assumption due to the presence of SAC. As such, spatial k-fold cross validation (S-KCV) [106] has been proposed to remove the SAC between the training and validation test set. Specifically, S-KCV implements a Euclidean 'dead-zone' area around all test points, such that all training points that lay in these areas are removed, see full definition in Chapter 6.

In Chapter 6, I put forward a newly improved road distance and travel time k-fold cross validation (RT-KCV) approach which proposes that non-Euclidean dead-zones better infer the spatial interactions of urban space. RT-KCV constructs *road network and travel time* dead-zones. The intuition for this method comes from both previous chapters 4 and 5. I show that RT-KCV outperforms the current state-of-the-art for estimating the generalisation performance of any geostatistical urban model across the interpolation-extrapolation range of application scenarios. As such, Chapter 6 presents a new method that can significantly improve the estimation of a model's generalisation performance in an urban setting.

7.2 Implications for Urban Science

The flow within cities is referred to as its metabolism [5] and the larger a city gets the more interrelated and diverse that metabolism becomes. A large and complex metabolism of diversity, networks and citizens can hence become a power in itself [105] which must be managed [9] to be understood and to ensure that our cities of the future are sustainable, efficient and promote a high quality of life [119].

Consequently, the research in this thesis provides a wealth of opportunities

for most, if not all, urban challenges. Such challenges could include houses price prediction, traffic flow estimation, understanding noise pollution intensity, planning emergency services, identifying the causes of traffic incidents, predicting the distribution of crime, optimising public transport routes, planning green space or modelling the density of air pollution.

The solutions put forth for each of these challenges are: (1) an improved spatial modelling procedure, compared with the current (Euclidean-based) state-of-the-art; (2) the opportunity to improve space as a feature of any urban hedonic model and (3) the ability to better estimate a model's generalisation performance compared with traditional KCV or S-KCV methods.

7.3 Implications for Geostatistics and Other Disciplines

The basic concept of Geostatistics is that variables of a specific geographic region tend to have a predictable structure. This domain is mostly discussed with respect to geology and mining [88]. Interpolation is the main tool within this subject area (i.e., Kriging) and uses Euclidean distance between observations to do so.

Despite this, random fields typically contain edges, breaks and other constraints and this applies to most environments, not just those which are urban. For example smog is blocked by hills and skyscrapers; animal migration can be restricted by lakes, mountains or settlements and contaminants can follow along a coastline or river. In addition, temperatures and precipitation can follow non-Euclidean patterns as well as magna records, gravel content, soil type and land use.

The solutions considered for each of these challenges are: (1) a newly defined procedure for mapping non-Euclidean and non-PD distances to a valid Euclidean metric for spatial modelling; (2) the opportunity to improve the measure of space as a feature of any hedonic model and (3) the ability to better estimate a

model's generalisation performance compared with traditional KCV or S-KCV methods.

These solutions are particularly important in the research areas of Geology and mining, where specialists typically already understand the interactions pertaining to a geological problem, and they simply use models to provide large scale estimates where manual data collection and expertise cannot be sustained. Hence, it may be common that the distance matrices that affect the interactions in, say mining or animal migrations, can be accurately defined and mapped to a valid metric. This is unlike urban systems where road distance and travel time are just assumed to measure mobility and there may be distances which can better measure this.

The same solutions and reasoning apply to other disciplines which rarely consider space, and even more rarely consider non-Euclidean distance to measure space. Such disciplines may include Statistics, Data Science and Machine Learning.

7.4 Implications for the UK Real Estate Industry

By 2030 investable real estate is expected to have grown by more than 55%; amounting to a UK residential market value of £9.145tn [93]. Consequently, owners of real estate, policy makers and everyday home buyers are reaching for technological innovations to drive sustainable, low-risk decisions in a now less local market [108]. As such, industry are looking for machine learning algorithms (similar to my automated valuation models (AVMs)) to reliably understand real estate trends over a large area where market behaviour may differ significantly.

The research presented in Publications 1 and 3 provide the (1) motivation and (2) application of a real estate valuation model for the whole of England and Wales. This model focusses on house prices at an individual level and is currently considered the state-of-the-art in the United Kingdom (UK) real estate

industry. A summary of the method is discussed below. A more comprehensive description is found in Publication 3.

The method, entitled 'SPENT', contains four-stages, as seen in figure 7.1 where the top left images represent stage 1, the middle images represent stage 2, the bottom represent stage 3 and the far right hand figure represents the fourth and final stage. The data utilised in this method is described fully in Section 3.2. Stage 1 produces a time singular dataset D^T from dataset D using the same approach taken in Chapters 4-6, that is, a space-time reduction using government defined output areas. Stage 2 undertakes a simple Euclidean-based Kriging interpolation on dataset D^T, obtaining an r^2 of 0.839 with 10-fold cross validation (CV). Stage 3 then introduces a set of property, network and economic features defined in Table 7.1. Finally, stage 4 puts all of the outputs from stages 1-3 into a single Gaussian process regression (GPR) [110], which produces an unprecedented r^2 of 0.966 and 0.920 with 10-fold and checkerboard CV respectively.

A market leading real estate tool named 'NimbusMaps', developed by Assured Property Group, embeds this technique. The interactive on-line tool allows a customer to select any property in England and Wales and in response, title ownership information, property size, flood risk and **estimated residential value** are provided. Consumers typically use this tool for site searching, site suitability analysis and market analysis.

This method is the current state-of-the-art in the industry, however we propose improvements with additional research from this thesis by: (1) introducing an estimated combined road distance and travel time metric in stage 2, such that the k-fold results of 0.839 can be improved and (2) providing an improved CV method. The primary benefit of this will be increased integrity, trust and confidence in the product. The new RT-KCV method allows a more accurate estimate of generalisation performance.

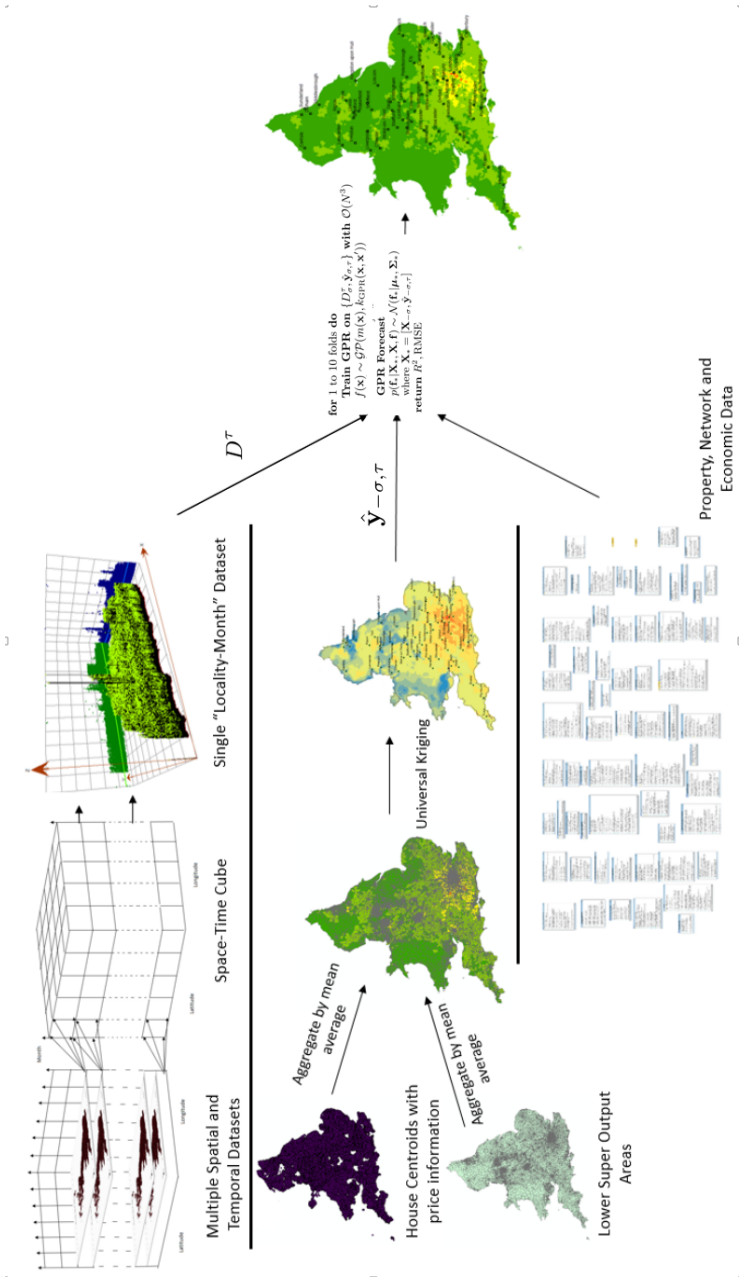


Figure 7.1: Process diagram corresponding to the space-property-economic-network-time (SPENT) algorithm.

Table 7.1: Property, network and economic features considered in my GPR AVM (entitled SPENT).

Category	Feature name	Description
Property	Footprints	Area of buildings
	Height	Average height of main building
	Size	Area of entire title
	Type	Detached, terraced, apartment etc.
	Tenure	Freehold or leasehold
	Status	New or old build
Network	Primary schools	Proximity and performance of closest
	Secondary schools	Proximity and performance of closest
	Train station	Proximity and usage
	Traffic flow	Passing the property
	Population	Postcode, 250, 500 and 1,000 meters.
Economic	Interest	Variable mortgage interest rate
	Sales rate	Total number of houses sold each month
	Inflation	Percentage value
	USD exchange rate	Ratio
...	...	...

7.5 Limitations to Generalisation

"What is the city but the people?"

William Shakespeare, The tragedy of Coriolanus.

In urban space, there will be observations whose SAC are not affected by road distance and travel time, for example, the height of citizens. In geostatistics and other disciplines, there will be applications where Euclidean distance does represent the best fit for models for example, the flight path of birds. Socially, there will be periods where trends cause interactions to act differently, for example the effect that political unrest can have on house prices. These limitations are discussed in more detail below.

Limitation 1. *Mobility across cities.* Residents of London, New York, Tokyo and other major international cities typically rely on public transport to move around, including trains and the underground. These modes of transportation are likely to partially determine a resident's perception of space around the city. This is compared with, say Birmingham or Coventry, where driving is the norm

[7]. This limitation provides a potential avenue for research, whereby one could consider the route and journey time on public transport. Interestingly, some city centres are becoming car-free, such as Birmingham, Chester, Oxford and Cambridge, all in the UK. This is likely to make a citizen's perception of space more complex again, especially given that these are new trends which could make historic data redundant or less relevant.

Limitation 2. *House anomolies.* There are some features in a property which may affect the price of a house (above and beyond its neighbours), for example the quality of the interior of the property, the foundations of the property, or whether the property sits at the bottom of a hill, where there may be a flood zone. In fact, in the latter example, a Euclidean distance may be more appropriate. Publication 3 includes land height and flood zones in its AVM, which does address this particular problem. It does however highlight that spatial interpolation alone is not appropriate.

Limitation 3. *Spatial generalisability.* All experiments in this thesis are tested on Coventry and the West Midlands. Other cities around the UK may in fact interact differently. For example, urban residents in one city may consider a 2 hour commute to work acceptable, whereas residents from another city may only accept a 30 minute commute. As such, the resident's perception of mobility can considerably differ from city to city [34]. This promotes the requirement for non-global Kriging methods to be modelled across multiple cities.

Limitation 4. *The trends of urban citizens.* Over time, the demographics of citizens change, for example, some UK cities are witnessing a large growth of older people [22], whilst other cities are attracting affluent young families [3]. These changes in the city bring different interactions, which alter a citizen's perception of the city, and hence the way observations, such as house prices relate spatially. For example, walking with a pushchair to the park versus catching

a bus to the local community centre. This thesis assumes that road distance and travel time are always the most appropriate measures of distance in a city, however demographics may dictate that other distance matrices are more appropriate for specific cities i.e., public transport, walking, the safest route etc.

CHAPTER 8

Conclusions and Further Work

8.1 Conclusions

This thesis has explored a number of issues arising from the requirement to sustainably manage elements of urban growth. Specifically, it answered three key questions that identify how to improve spatial models in urban areas. These were: (1) which distance function best models real world spatial interactions in an urban setting? (2) when, if ever, are non-Euclidean distance functions valid for urban spatial models? and (3) what is the best way to estimate the generalisation performance of urban spatial models? Each question was studied by means of three contributions.

Contribution 1 considered RQ1 by proposing three approximate restricted road distance, travel time and combined pairwise distance matrices to inform spatial autocorrelation (SAC). The estimated value was the Minkowski distance function most correlated to the OpenStreetMaps (OSM) road network data. This provided valid distance metrics for spatial interpolation. The work in this contribution was taken from Publication 4 in Section 1.2.

the second contribution addressed RQ2 by proposing a method to approximate restricted road distance, journey time and combined matrices using an embedded lower-dimensional Euclidean space. This method ensured that covariance and (semi)variance functions remained valid for spatial interpolation when using urban-specific non-Euclidean distances. The work in this contribution was taken from Publication 5 in Section 1.2.

In Chapter 6, RQ3 is addressed by introducing a new spatial k-fold cross validation method; road distance and travel time k-fold cross validation (RT-KCV).

This method constructed *road network and travel time* dead-zones to better estimate and remove urban SAC. I showed that RT-KCV outperforms the current state-of-the-art for estimating the generalisation performance of any geostatistical urban model across the interpolation-extrapolation range of application scenarios. The work in this contribution was taken from Publication 6 in Section 1.2.

In Chapter 7, I discuss the implications of my research for urban science, geostatistics and the UK real estate industry. This chapter also considered the impact that my work will have on an existing nationwide house price predictor, which is discussed in publications 1 and 3 and is considered the current state-of-the-art in the industry.

Throughout, all experiments utilised real-world datasets in England and Wales, most notably: restricted roads, travel times, house sales and traffic counts. With these datasets, I displayed a set of case studies which showed the potential to improve model accuracy by 2 times against Euclidean distances and, in some cases, a 90% improvement for the estimation of generalisation performance.

Combined, the contributions improved the way that proximity-based urban models perform and also provided a more accurate estimate of generalisation performance for predictive models in urban space.

8.2 Recommendations for Future Research

The opportunities that have been opened up by the research in this thesis are substantial. Contributions **1** and **2** prove that techniques from disciplines outside of spatial statistics can massively improve real world applications of spatial models in restricted space and Contribution **3** shows that the generalisation performance of geographically-dependant extrapolation techniques have been misinformed for years. As such, the following suggestions provide a set of potential routes that could be taken to further improve an abundance of scientific

applications as a result of the work in this thesis. The suggestions also offers a recommended direction for making significant theoretical improvements to spatial data analysis, machine learning (ML) and artificial intelligence (AI).

1. One could use the techniques presented in this thesis for other urban science problems i.e., estimating the intensity of noise or air pollution, the causes of traffic incidents, the distribution of crime or the optimal location of urban green space. These applications could benefit from using road distance and travel time to inform (1) spatial models and (2) the approximation of generalisation performance to any model containing spatial data.

2. In addition, one could administer each experiment with alternative models such as Universal Kriging or geographically weighted regression (GWR). In testing these models, it may be possible to further improve my results. It is important to note that the findings in Contribution **3** do not rely on spatial models, but instead data that contain SAC only.

3. One could also lead an experiment to test my isometric embedding approach on other non-urban, non-Euclidean, non-positive definite (PD) distance matrices for non-urban problems. For example, animal migration, rainfall, soil type and land use.

4. Furthermore, one could conduct an experiment to better understand how to combine road distance and travel time for Kriging optimisation. This could allow for a more confident estimation of a person's perception of space.

5. Finally, one could coordinate an experiment to optimise the metric used to predict the operating point for my RT-KCV estimate in Contribution **3**. The operating point determines the success of a cross validation technique at estimating the generalisation performance of a model, and hence if the operating point is improved, the RT-KCV technique can too be enhanced.

Bibliography

[1] W. Alonso. A theory of the urban land market. *Papers in Regional Science*, 6(1):149–157, 1960.

[2] L. Anselin. Local indicators of spatial association—lisa. *Geographical analysis*, 27(2):93–115, 1995.

[3] R. Atkinson. Padding the bunker: strategies of middle-class disaffiliation and colonisation in the city. *Urban Studies*, 43(4):819–832, 2006.

[4] Axiom. Personicx cluster perspectives; lifestyle, digital and financials. *http://www.acxiom.com/personicx/*, 2014. Accessed: 2016-09-01.

[5] P. Baccini. A city's metabolism: Towards the sustainable development of urban systems. *Journal of Urban Technology*, 4(2):27–39, 1997.

[6] S. Banerjee. On geodetic distance computations in spatial modeling. *Biometrics*, 61(2):617–625, 2005.

[7] D. Banister, S. Watson, and C. Wood. Sustainable cities: transport, energy, and urban form. *Environment and Planning B: planning and design*, 24(1):125–143, 1997.

[8] S. Basu and T. Thibodeau. Analysis of spatial autocorrelation in house prices. *The Journal of Real Estate Finance and Economics. 17: 61.*, 1998.

[9] M. Batty. *Cities and Complexity: Understanding Cities with Cellular Automata, Agent-Based Models, and Fractals*. The MIT Press, 2007. ISBN: 0262524791.

[10] M. Batty. *The New Science of Cities*. The MIT Press, 2013. ISBN: 0262019523.

[11] M. Batty, K. W. Axhausen, F. Giannotti, A. Pozdnoukhov, A. Bazzani, M. Wachowicz, G. Ouzounis, and Y. Portugali. Smart cities of the future. *The European Physical Journal Special Topics*, 214(1):481–518, 2012.

[12] D. Biello. Gigalopolises: Urban land area may triple by 2030. *https://www.scientificamerican.com/article/cities-may-triple-in-size-by-2030/*, 2012.

[13] S. Bond, S. Sims, and P. Dent. *Towers, Turbines and Transmission Lines: Impacts on Property Value*. John Wiley & Sons, 2013.

[14] A. Can. The measurement of neighborhood dynamics in urban house prices. *Economic geography*, 66(3):254–272, 1990.

[15] A. Caplin, S. Chopra, J. V. Leahy, Y. LeCun, and T. Thampy. Machine learning and the spatial structure of house prices and housing returns. *Available at SSRN 1316046*, 2008.

[16] K. Case and R. Shiller. Prices of single-family homes since 1970: new indexes for four cities. *New England Economic Review. Sept./Oct. 45-56*, 1987.

[17] D. Castelvecchi. Can we open the black box of ai? *Nature News*, 538(7623):20, 2016.

[18] T. Champion. People in cities: the numbers. *Future of cities: working paper. Foresight, Government Office for Science*, 2014.

[19] W. G. G. Changling. Application of the kriging technique in geography [j]. *Acta Geographica Sinica*, 4:008, 1987.

[20] J. Chica Olmo. Spatial estimation of housing prices and locational rents. *Urban studies*, 32(8):1331–1344, 1995.

[21] W. Christaller. *Die Zentralen Orte in Suddeutscland*. Prentice Hall, Englewood Cliffs, NJ., 1933.

[22] K. Christensen, G. Doblhammer, R. Rau, and J. W. Vaupel. Ageing populations: the challenges ahead. *The lancet*, 374(9696):1196–1208, 2009.

[23] J. M. Clapp, M. Rodriguez, and G. Thrall. How gis can put urban economic analysis on the map. *Journal of Housing Economics*, 6(4):368–386, 1997.

[24] A. Cliff and K. Ord. Testing for spatial autocorrelation among regression residuals. *Geographical analysis*, 4(3):267–284, 1972.

[25] A. D. Cliff and J. K. Ord. *The problem of spatial autocorrelation*. University of Bristol, Department of Economics and Department of Geography, 1968.

[26] A. D. Cliff and J. K. Ord. Spatial and temporal analysis: autocorrelation in space and time. *Quantitative geography: a British view*, pages 104–110, 1981.

[27] R. H. Coase. Economics and contiguous disciplines. In *The organization and retrieval of economic knowledge*, pages 481–495. Springer, 1977.

[28] D. R. Cox and H. D. Miller. *The theory of stochastic processes*. Routledge, New York, 2017. ISBN: 9781351408950.

[29] C. A. G. Crawford and L. J. Young. Geostatistics: what's hot, what's not, and other food for thought. In *Proceedings of the 8th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Shanghai, PR China*, pages 8–16, 2008.

[30] N. Cressie. Fitting variogram models by weighted least squares. *Journal of the International Association for Mathematical Geology*, 17(5):563–586, 1985.

[31] N. Cressie. Spatial prediction and ordinary kriging. *Mathematical Geology*, 20(4):405–421, May 1988.

[32] N. Cressie. The origins of kriging. *Mathematical geology*, 22(3):239–252, 1990.

[33] N. Cressie. Statistics for spatial data. *Terra Nova*, 4(5):613–617, 1992.

[34] T. Cresswell. *Place: an introduction*. John Wiley & Sons, 2014. ISBN: 9780470655627.

[35] H. Crosby, T. Damoulas, A. Caton, P. Davis, J. Porto de Albuquerque, and S. A. Jarvis. Road distance and travel time for an improved house price kriging predictor. *Geo-spatial Information Science*, 21(03):185–194, 2018.

[36] H. Crosby, T. Damoulas, and S. A. Jarvis. Embedding road networks and travel time into distance metrics for urban modelling. *International Journal of Geographic Information Sciences (IJGIS)*, Awaiting Editing, 2018.

[37] H. Crosby, P. Davis, T. Damoulas, and S. Jarvis. A spatiotemporal, gaussian process regression, real-estate price predictor. *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 68:1–68:4, article No. 68, DOI: 10.1145/2996913.2996960*, 2016.

[38] H. Crosby, P. Davis, and S. A. Jarvis. Spatially-intensive decision tree prediction of traffic flow across the entire uk road network. *Distributed Simulations and Real Time Applications (DS-RT16) London, UK, September 21-23*, 2016.

[39] F. C. Curriero. On the use of non-euclidean distance measures in geostatistics. *Mathematical Geology*, 38(8):907–926, 2006.

[40] L. Daniel. Gis helping to reengineer real estate. volume 3. Earth Observation Magazine, 1994.

[41] DepartmentForTransport. Road lengths in great britain 2016. *UK Government (Crown copyright)*, 2017.

[42] K. Donlon. Using gis to improve the services of a real estate company. *Department of Resource Analysis, Saint Mary's University of Minnesota, Winona, MN. 59987*, 2007.

[43] R. A. Dubin. Spatial autocorrelation and neighborhood quality. *Regional science and urban economics*, 22(3):433–452, 1992.

[44] R. A. Dubin. Predicting house prices using multiple listings data. *The Journal of Real Estate Finance and Economics*, 17(1):35–59, 1998.

[45] J. P. Elhorst, M. M. Fischer, and A. Getis. Handbook of applied spatial analysis. *Methods*, pages 377–407, 2010.

[46] A. Elisseeff, M. Pontil, et al. Leave-one-out error and stability of learning algorithms with applications. *NATO science series sub series iii computer and systems sciences*, 190:111–130, 2003.

[47] J. K. Eom, M. S. Park, T.-Y. Heo, and L. F. Huntsinger. Improving the prediction of annual average daily traffic for nonfreeway facilities by applying a spatial statistical method. *Transportation Research Record*, 1968(1):20–29, 2006.

[48] G. Evans. Creative cities, creative spaces and urban policy. *Urban studies*, 46(5-6):1003–1040, 2009.

[49] A. S. Fotheringham. Scale-independent spatial analysis. *Accuracy of spatial databases*, pages 221–228, 1989.

[50] S. Fotheringham and P. Rogerson. *Spatial analysis and GIS*. CRC Press, 2013. ISBN: 0718401032.

[51] L. M. Ganio, C. E. Torgersen, and R. E. Gresswell. A geostatistical approach for describing spatial pattern in stream networks. *Frontiers in Ecology and the Environment*, 3(3):138–144, 2005.

[52] J. Gantz and D. Reinsel. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. 2012.

[53] P. H. Garcia-Soidan, M. Febrero-Bande, and W. Gonzalez-Manteiga. Nonparametric kernel estimation of an isotropic variogram. *Journal of Statistical Planning and Inference*, 121(1):65–92, 2004.

[54] D. Gayle. Daily commute of two hours – reality for 3.7m uk workers. *The Guardian UK. Work Life Balance.*, 2017.

[55] A. Getis. A history of the concept of spatial autocorrelation: A geographer's perspective. *Geographical Analysis*, 40(3):297–309, 2008.

[56] A. Getis and J. K. Ord. The analysis of spatial association by use of distance statistics. *Geographical analysis*, 24(3):189–206, 1992.

[57] A. Getis and J. K. Ord. Local spatial statistics: an overview. *Spatial analysis: modelling in a GIS environment*, 374:261–277, 1996.

[58] M. Goodchild and L. Li. Formalizing space and place. In *CIST2011-Fonder les sciences du territoire*, pages 177–183, 2011.

[59] E. Griffin. *A short history of the British industrial revolution.* Macmillan International Higher Education, 2010. ISBN: 13:9780230579255.

[60] D. K. Hayunga and A. Kolovos. Geostatistical space–time mapping of house prices using bayesian maximum entropy. *International Journal of Geographical Information Science*, 30(12):2339–2354, 2016.

[61] M. Helbich, W. Brunauer, E. Vaz, and P. Nijkamp. Spatial heterogeneity in hedonic house price models: The case of austria. *Urban Studies*, 51(2):390–411, 2014.

[62] N. Henretty. Housing affordability in england and wales: 2017. *Office of National Statistics*, 2018. Accessed 01 September 2018. https://www.ons.gov.uk/peoplepopulationandcommunity/housing/bulletins/housingaffordabilityinenglandandwales/2017.

[63] J. M. V. Hoef, E. Peterson, and D. Theobald. Spatial statistical models that use flow and stream distance. *Environmental and Ecological Statistics*, 13(4):449–464, Dec 2006.

[64] V. Houlden, S. Weich, and S. Jarvis. A cross-sectional analysis of green space prevalence and mental wellbeing in england. *BMC Public Health*, 17(1):460, May 2017.

[65] B. Huang, B. Wu, and M. Barry. Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. *International Journal of Geographical Information Science*, 24(3):383–401, 2010.

[66] L. J. Hubert and R. G. Golledge. A heuristic method for the comparison of related structures. *Journal of mathematical psychology*, 23(3):214–226, 1981.

[67] G. Hudson and H. Wackernagel. Mapping temperature using kriging with external drift: theory and an example from scotland. *International journal of Climatology*, 14(1):77–91, 1994.

[68] A. M. Ibrahim and B. Bennett. The assessment of machine learning model performance for predicting alluvial deposits distribution. *Procedia Computer Science*, 36:637–642, 2014.

[69] H.-Y. Kim. A geostatistical approach for improved prediction of traffic volume in urban area. *Journal of the Korean Association of Geographic Information Studies*, 13(4):138–147, 2010.

[70] L. J. King et al. Central place theory. *Regional Research Institute, West Virginia University Book Chapters*, pages 1–52, 1985.

[71] P. Knox and S. Pinch. *Urban social geography: an introduction*. Routledge, 2014. ISBN: 9781317903260.

[72] R. Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[73] Kok, Nils, Monkkonen, and Quigley. Economic geography, jobs, and regulations: the value of land and housing. *AREUEA Meetings Denver*, 2011.

[74] K. Krivoruchko and A. Gribov. Geostatistical interpolation and simulation in the presence of barriers. *geoENV IV—Geostatistics for Environmental Applications*, pages 331–342, 2004.

[75] M. Kuntz and M. Helbich. Geostatistical mapping of real estate prices: an empirical comparison of kriging and cokriging. *International Journal of Geographical Information Science*, 28(9):1904–1921, 2014.

[76] W. E. Larimore and R. K. Mehra. Problem of overfitting data. *Byte*, 10(10):167–178, 1985.

[77] R. Lark. Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood. *Geoderma*, 105(1-2):49–80, 2002.

[78] K. Le Rest, D. Pinaud, P. Monestiez, J. Chadoeuf, and V. Bretagnolle. Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Global ecology and biogeography*, 23(7):811–820, 2014.

[79] P. Legendre. Spatial autocorrelation: trouble or new paradigm? *Ecology*, 74(6):1659–1673, 1993.

[80] T. Liebig, N. Piatkowski, C. Bockermann, and K. Morik. Predictive trip planning-smart routing in smart cities. In *EDBT/ICDT Workshops*, pages 331–338, 2014.

[81] L. S. Little, D. Edwards, and D. E. Porter. Kriging in estuaries: as the crow flies, or as the fish swims? *Journal of experimental marine biology and ecology*, 213(1):1–11, 1997.

[82] B. Lu, M. Charlton, C. Brunsdon, and P. Harris. The minkowski approach for choosing the distance metric in geographically weighted regression. *International Journal of Geographical Information Science*, 30(2):351–368, 2016.

[83] B. Lu, M. Charlton, P. Harris, and A. S. Fotheringham. Geographically weighted regression with a non-euclidean distance metric: a case study using hedonic house price data. *International Journal of Geographical Information Science*, 28(4):660–681, 2014.

[84] Y. Lv, Y. Duan, W. Kang, Z. Li, and F. Y. Wang. Traffic flow prediction with big data: A deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):865–873, April 2015.

[85] Machin and Gibbons. Valuing school quality, better transport, and lower crime: evidence from house prices. *oxford review of Economic Policy 24.1 (2008): 99-119*, 2003.

[86] J. Malczewski. Gis-based land-use suitability analysis: a critical overview. *Progress in planning*, 62(1):3–65, 2004.

[87] A. Marshall. *Elements of economics of industry*, volume 1. Macmillan, International Journal of Ethics 3 (2):266-267, 1892.

[88] G. Matheron. Principles of geostatistics. *Economic geology*, 58(8):1246–1266, 1963.

[89] A. Matkan, A. Shakiba, B. Mirbagheri, and H. Tavoosi. A comparison between kriging, cokriging and geographically weighted regression models for estimating rainfall over north west of iran. In *10th EMS Annual Meeting, 10th European Conference on Applications of Meteorology (ECAM) Abstracts, held Sept. 13-17, 2010 in Zürich, Switzerland. http://meetings. copernicus. org/ems2010/, id. EMS2010-325*, 2010.

[90] McClusky and Borst. Specifying the effect of location in multivariate valuation models for residential properties. *Property Management,25, 312343*, 2007.

[91] Q. Meng. Regression kriging versus geographically weighted regression for spatial interpolation. *International journal of advanced remote sensing and GIS*, 3(1):pp–606, 2014.

[92] H. J. Miller. Tobler's first law and spatial analysis. *Annals of the Association of American Geographers*, 94(2):284–289, 2004.

[93] Mitchell. The size and structure of the uk property market: End-2016 update. http://www.ipf.org.uk/resourceLibrary/, 2017. Accessed June 2018.

[94] H. Miura. A study of travel time prediction using universal kriging. *TOP*, 18(1):257–270, Jul 2010.

[95] A. W. Moore and M. S. Lee. Efficient algorithms for minimizing cross validation error. In *Machine Learning Proceedings 1994*, pages 190–198. Elsevier, 1994.

[96] P. A. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950.

[97] R. R. Murphy, E. Perlman, W. P. Ball, and F. C. Curriero. Water-distance-based kriging in chesapeake bay. *Journal of Hydrologic Engineering*, 20(9):05014034, 2014.

[98] U. Nations. The worlds city's in 2016. World Urbanization Prospects Report ISBN 978-92-1-151549-7, Department of Economic and Social Affairs, United Nations, 2016.

[99] A. C. Nelson. Transit stations and commercial property values: a case study with policy and land-use implications. *Journal of Public Transportation*, 2(3), 1999.

[100] ONS. Output areas - an introduction to output areas - the building block of census geography. https://www.ons.gov.uk/census/2001censusandearlier/dataandproducts/outputgeography/outputareas, Published 2001. Accessed September 2018.

[101] OpenStreetMap Contributors. Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18, Oct. 2008.

[102] R. K. Pace, R. Barry, J. M. Clapp, and M. Rodriquez. Spatiotemporal autoregressive models of neighborhood effects. *The Journal of Real Estate Finance and Economics*, 17(1):15–33, 1998.

[103] R. K. Pace, R. Barry, and C. F. Sirmans. Spatial statistics and real estate. *The Journal of Real Estate Finance and Economics*, 17(1):5–13, 1998.

[104] R. K. Pace and O. W. Gilley. Using the spatial configuration of the data to improve estimation. *The Journal of Real Estate Finance and Economics*, 14(3):333–340, 1997.

[105] S. Pincetl, P. Bunje, and T. Holmes. An expanded urban metabolism method: Toward a systems approach for assessing urban energy processes and causes. *Landscape and urban planning*, 107(3):193–202, 2012.

[106] J. Pohjankukka, T. Pahikkala, P. Nevalainen, and J. Heikkonen. Estimating the prediction performance of spatial models via spatial k-fold cross validation. *International Journal of Geographical Information Science*, 31(10):2001–2019, 2017.

[107] J. Pohjankukka, T. Pahikkala, P. Nevalainen, and J. Heikkonen. Estimating the prediction performance of spatial models via spatial k-fold cross validation. *International Journal of Geographical Information Science*, 31(10):2001–2019, 2017.

[108] PwC. Real estate 2020 building the future. `https://www.pwc.com/gx/en/industries/financial-services/asset-management/publications/real-estate-2020-building-the-future.html`, 2017.

[109] G. Randall. The effects of subdivision design on housing values: the case of alleyways. *Journal of Real Estate Research*, 23(3):265–274, 2002.

[110] C. E. Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.

[111] D. R. Roberts, V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 2017.

[112] E. Rothschild. Adam smith and the invisible hand. *The American Economic Review*, 84(2):319–322, 1994.

[113] A. Se Can and I. Megbolugbe. Spatial dependence and house price index construction. *The Journal of Real Estate Finance and Economics*, 14(1-2):203–222, 1997.

[114] K. C. Seto, B. Güneralp, and L. R. Hutyra. Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools.

Proceedings of the National Academy of Sciences, 109(40):16083–16088, 2012.

[115] R. Shahid, S. Bertazzon, M. L. Knudtson, and W. A. Ghali. Comparison of distance measures in spatial analytical modeling for health service planning. *BMC Health Services Research*, 9(1):200, Nov 2009.

[116] S. Stewart, K. MacIntyre, S. Capewell, and J. McMurray. Heart failure and the aging population: an increasing burden in the 21st century? *Heart*, 89(1):49–53, 2003.

[117] M. Stone. Cross-validation and multinomial prediction. *Biometrika*, 61(3):509–515, 1974.

[118] S. Sun, C. Zhang, and G. Yu. A bayesian network approach to traffic flow forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 7(1):124–132, March 2006.

[119] J. Swanson. *Designing the Urban Future: Smart Cities*. Scienctific American, 75 Varick Street, 9th Floor, New York, 978-1-466842618, 2014.

[120] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[121] R. Thaler. A note on the value of crime control: evidence from the property market. *Journal of Urban Economics*, 5(1):137–145, 1978.

[122] P. G. Theodoridou, G. P. Karatzas, E. A. Varouchakis, and G. A. Corzo Perez. Geostatistical analysis of groundwater level using euclidean and non-euclidean distance metrics and variable variogram fitting criteria. In *EGU General Assembly Conference Abstracts*, volume 17, 2015.

[123] J. Thunen. *Der Isolierte Staat in Beziehung auf Landwirtschaft und Nationalkonomie, Hamburg, Perthes. English translation by C.M. Wartenberg*. Oxford University Press, 1826.

[124] A. Townsend. Cities of data: Examining the new urban science. *Public Culture*, 27(2 (76)):201–212, 2015.

[125] H. Wackernagel. *Ordinary Kriging*, pages 74–81. Springer Berlin Heidelberg, Berlin, Heidelberg, 1995.

[126] J. Wang, P. Shang, and X. Zhao. A new traffic speed forecasting method based on bi-pattern recognition. *Fluctuation and Noise Letters*, 10(01):59–75, 2011.

[127] X. Wang and K. M. Kockelman. Forecasting network data: Spatial interpolation of traffic counts from texas data. *Transportation Research Record*, 2105(1):100–108, 2009.

[128] W. C. Wheaton. A comparative static analysis of urban spatial structure. *Journal of Economic Theory*, 9(2):223–237, 1974.

[129] B. M. Williams and L. A. Hoel. Modeling and forecasting vehicular traffic flow as a seasonal arima process: Theoretical basis and empirical results. *Journal of transportation engineering*, 129(6):664–672, 2003.

[130] Y.-J. Wu, F. Chen, C. Lu, B. Smith, and Y. Chen. Traffic flow prediction for urban network using spatio-temporal random effects model. In *91st Annual Meeting of the Transportation Research Board (TRB)*, 2012.

[131] H. Yin, S. Wong, J. Xu, and C. Wong. Urban traffic flow prediction using a fuzzy-neural approach. *Transportation Research Part C: Emerging Technologies*, 10(2):85–98, 2002.

[132] G. Yu, J. Hu, C. Zhang, L. Zhuang, and J. Song. Short-term traffic flow forecasting based on markov chain model. In *Intelligent Vehicles Symposium, 2003. Proceedings. IEEE*, pages 208–212. IEEE, 2003.

[133] K. Yu, J. Mateu, and E. Porcu. A kernel-based method for nonparametric estimation of variograms. *Statistica Neerlandica*, 61(2):173–197, 2007.

[134] R. Zas. Iterative kriging for removing spatial autocorrelation in analysis of forest genetic trials. *Tree genetics & genomes*, 2(4):177–185, 2006.

[135] D. W. Zimmerman and B. D. Zumbo. Rank transformations and the power of the student t test and welch t'test for non-normal populations with unequal variances. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 47(3):523, 1993.

[136] H. Zou, Y. Yue, Q. Li, and A. G. Yeh. An improved distance metric for the interpolation of link-based traffic data using kriging: a case study of a large-scale urban road network. *International Journal of Geographical Information Science*, 26(4):667–689, 2012.