ARTICLE

Road and Travel Time Cross-Validation for Urban Modelling

Henry Crosby[a,b,+*] and Theodore Damoulas[b,c,+] and Stephen A. Jarvis[a,b,+]

[a] *Warwick Institute for the Science of Cities*; [b] *Dept. of computer Science*; [c] *Dept. of Statistics*; [+] *University of Warwick, UK*

(*Received 00 Month 200x; final version received 00 Month 200x*)

The physical and social processes in urban systems are inherently spatial and hence data describing them contain spatial autocorrelation (a proximity-based interdependency on a variable) that need to be accounted for. Standard k-fold cross-validation (KCV) techniques that attempt to measure generalisation performance of machine learning and statistical algorithms are inappropriate in this setting due to their inherent i.i.d assumption, which is violated by spatial dependency. As such, more appropriate validation methods have been considered, notably blocking (Roberts *et al.* 2017) and spatial k-fold cross-validation (SKCV) (Pohjankukka *et al.* 2017). However, the physical barriers and complex network structures which make up a city's landscape means that these methods are also inappropriate, largely because the travel patterns (and hence Spatial Autocorrelation (SAC)) in most urban spaces are rarely Euclidean in nature. To overcome this problem, we propose a new road distance and travel time k-fold cross-validation method, RT-KCV. We show how this outperforms the prior art in providing better estimates of the true generalisation performance to unseen data.

Keywords: road distance, travel time, cross-validation, urban science, GIS.

1. Introduction

Cities currently accommodate 55% of the worlds population. By 2050 the urban population is estimated to grow by 2.5 billion people, resulting in a 13% increase on today's figure (Prudhomme 2018). Consequently, policy makers are looking to utilise the vastly increasing amounts of urban data to build smarter models for sustainable and high quality urban life. Given this, an increasing number of spatial models are being

*Corresponding author. Email: h.j.crosby@warwick.ac.uk

built in the domain of urban sciences, e.g. (Zou *et al.* 2012) and (Crosby *et al.* 2016). In order to correctly estimate how well these models generalise to unseen locations, cross-validation is commonly used, which itself must account for inherent dependency structures in spatiotemporal datasets.

Proximity-based urban interactions result in data which are not independent and identically distributed (i.i.d). Various techniques such as semi-variograms (Cressie 2015), Moran's I (Moran 1950) or Getis's G (Getis and Ord 1992) have been proposed to statistically measure the extent of these dependencies. Additionally, spatial k-fold cross-validation (SKCV) (Pohjankukka *et al.* 2017), blocking (Roberts *et al.* 2017) and stratified sampling (Crosby *et al.* 2016) consider the effect of these spatial dependencies by accounting for the Spatial Autocorrelation (SAC) between test and training points, which traditional cross-validation (CV) methods do not. Specifically, SKCV 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. However, we argue that Euclidean distances may not be appropriate for urban systems.

In this paper, we introduce a new spatial k-fold validation method, termed 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 (SKCV (Pohjankukka *et al.* 2017)) and (3) improved performance in capturing and removing urban SAC. We demonstrate these contributions across two large-scale urban datasets and three different scenarios of interpolation-extrapolation. We also provide an extensive experimental comparison across multiple CV techniques and offer a systematic way to choose the dead-zone distance. We argue that RT-KCV should be the spatial-validation method of choice for urban modelling problems.

1.1. *Paper Overview*

Section 2 provides a full description of why and under what settings SAC removal is required. Section 3 reviews related approaches for SAC detection and generalisation performance (i.e., a model's ability to generalise to an unseen location). Thereafter, Section 4 redefines spatial cross-validation for urban spaces, utilising a unique set of restricted road, travel-time and combined distance dead-zones. Section 5 introduces two urban datasets over three modelling settings with the purpose of comparing the estimated generalisation performance of several validation methods KCV, SKCV, R-KCV, T-KCV, RT-KCV and blocking KCV. Finally, Section 6 concludes our findings and puts forward possible avenues for future research.

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 (Stone 1974). The main purpose of cross-validation is to estimate how well a model will generalise to unseen data, sometimes referred to as a **ground truth test set**.

As such, cross-validation is also used to detect overfitted models. Specifically, k-fold cross-validation (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 similar SAC 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, traditional i.i.d is typically assumed in cross-validation, which is over-optimistic (i.e., overestimates the generalisation performance of the model), unless the model is attempting pure interpolation. In this manuscript, 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 (Getis 1995).

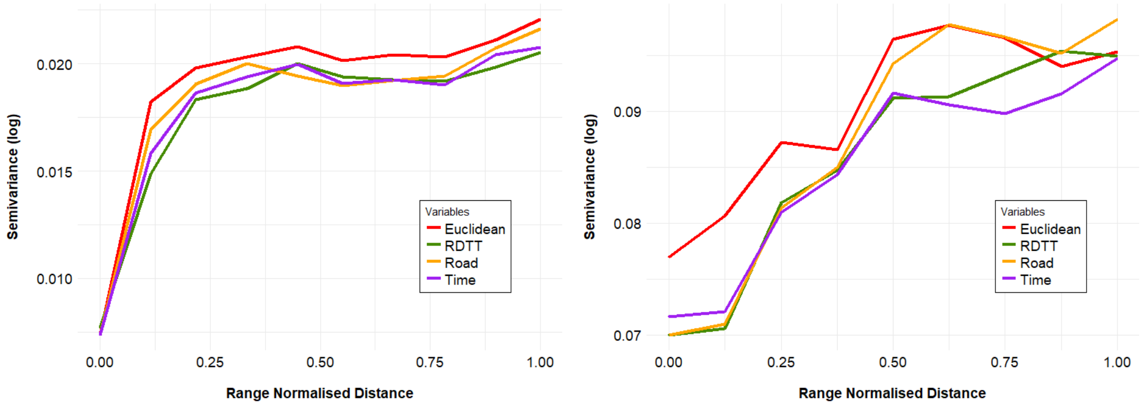
Removing $SAC_{traintest}$ to improve the estimate of a model's generalisation performance requires an understanding of how dependencies are structured and unfold in geographic space. Typically it is assumed that spatial dependence is Euclidean in nature, but in most urban settings natural or man-made restrictions (e.g., one-way road systems) violate this assumption (Cressie 2015, Crosby *et al.* 2018). As such, we hypothesise that road distance, travel time and a combination of both are better able to infer urban SAC.

2.0.1. Motivating our Hypothesis.

Figures 1(a)-(b) utilise Euclidean, road distance (RD), travel time (TT) and RDTT (a combination of road distance and travel time, as defined in Section 4) to provide semivariograms for two urban case studies (real estate valuation and traffic flow). It can be seen that the semivariance of each pairwise Euclidean distance is larger than the semivariance of our proposed distance functions. Hence, the correlation contained at each Euclidean distance lag is smaller, i.e., SAC drops faster when proximity in Euclidean space gets larger in comparison with our proposed non-Euclidean distances. Notably, road distance and travel time are more efficient at discovering and capturing SAC at all lags for both case studies.

2.0.2. 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, a combined road distance and travel time matrix (RDTT) is calculated as it affords the opportunity to take into account



(a) House Prices (b) Traffic Flow

Figure 1.: Range normalised semivariograms for two case studies (house prices and traffic flows) with four distance measures (Euclidean, road distance, travel time and combined).

the exclusive behaviors of both matrices, this is why, we believe, our RT-KCV method is shown to outperform all studied approaches in our case studies found in Section 5.

3. Related Literature

3.1. *Spatial Autocorrelation (SAC)*

SAC describes correlation between all observed variables to each other in a spatial dataset. This correlation can be explained solely by geographical proximity (Hubert and Golledge 1981). SAC was first influenced by the central place theory (Christaller 1933), which in itself was inspired by theories of proximity and nearness (Thunen 1826). Later, SAC and Moran's I were developed (Cliff and Ord 1968, Moran 1950), and gave rise to a series of measures, such as Getis and Ord's Gi statistic (Getis and Ord 1996) and Matheron's $1/\gamma$ (inverse of the semivariogram) (Matheron 1963).

Commonly, SAC is measured to (i) test model mis-specifications (Cliff and Ord 1972), (ii) measure the strength of spatial effects on a variable, (iii) test for spatial stationarity, heterogeneity or clustering, (iv) detect distance decay (v) identify outliers and (vi) design spatial samples (Getis 2008, Anselin 1995, Fotheringham and Rogerson 2013). In this work, we remove SAC between training and validation test sets in order to better estimate the generalisation of our models in different settings. With the exception of research by Zas (2006) and Ibrahim and Bennett (2014) whose experiments both utilise blocking (see section 3.2.1 for more details on this method), little research has considered the different sources of SAC (SAC_{train}, SAC_{test} and $SAC_{traintest}$). No research has utilised non-Euclidean SAC to improve the validation process and hence our estimates of model generalisation.

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.

3.2.1. 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, which is inappropriate in applications of data containing spatial, temporal, grouping and hierarchical autocorrelation (Roberts *et al.* 2017). 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 (Crosby *et al.* 2016) 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 versus extrapolation) and (3) contains SAC at each strata border.

3.2.2. 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 (Kohavi *et al.* 1995). The most typical cases of k-fold cross-validation are k=10 and k=n (Leave-One-Out), where n is the total number of points and hence the latter model trains on the largest set of data possible, but is time consuming on large datasets (Elisseeff *et al.* 2003). Traditional KCV withholds the central independence assumption (i.i.d) which, as discussed, can provide optimistic estimates of generalisation performance (Larimore and Mehra 1985, Roberts *et al.* 2017).

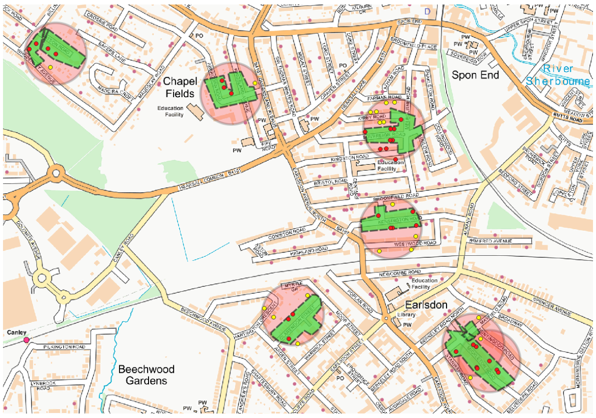
As such, Geostatisticians are critical of cross-validation for confirmatory data analysis with dependent data (Cressie 1990). Spatially aware cross-validation methods have hence been proposed to break the dependence between the training and testing set. The most notable of these methods is spatial k-fold cross-validation (SKCV), which estimates a predictor's performance by first implementing traditional k-fold cross-validation and secondly, removing all training points within an empirically designed Euclidean dead-zone from all test points (Pohjankukka *et al.* 2017). Additionally, (Le Rest *et al.* 2014) proposes a special case of SKCV termed spatial leave-one-out (SLOO), which computes a threshold distance equal to the range of residual SAC 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 an optimal dead-zone radii (instead, SKCV establishes how well the method captures the model's actual prediction capability for a specific user-defined dead-zone); (3) the lack of specificity toward the 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, although SKCV is not limited to Euclidean, the author only shows examples containing it. As we will show, our 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 (Moore and Lee 1994). The benefits of this approach, over k-fold cross-validation techniques, are its ability to simulate unseen values in unseen areas. 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 (Crosby *et al.* 2018) or irregularly arranged regular or irregular blocks (Roberts *et al.* 2017). 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 high amount of SAC present at block borders and (4) ad-hoc choices for the shape, size, placement and regularity of the blocks. As we demonstrate, our proposed RT-KCV method overcomes all of these issues.

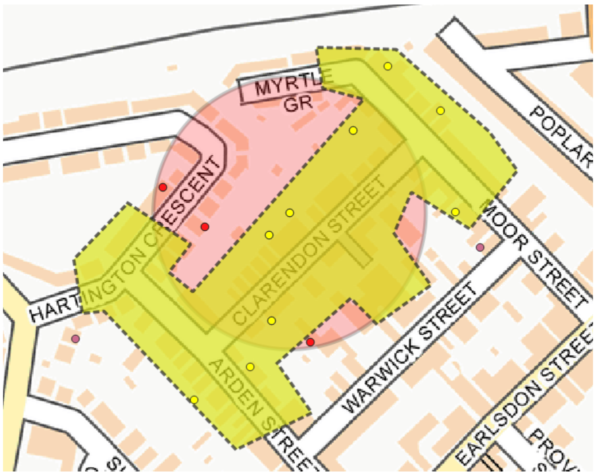
4. Road and Travel Time Validation

RT-KCV is a spatial dead-zone technique which, in a similar way to SKCV, constructs an area around each test point from which all training points are removed. Unlike the current examples utilising SKCV, 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 SAC in access-restricted areas such as cities. Figure 2 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. See Section 5.2.2 for an in-depth description of how the convex hull of the isochrone is computed. Figure 2(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. We show that this is the case in two real world urban datasets and conjecture that this generalises to a plethora of urban applications driven by human behavior, such as evaluating the impact of green space, designing algorithms for car sharing, predicting house prices and designing methods to improve traffic flow. The definition below describes the entire process of the combined road distance and travel time RT-KCV method, which is compared against (1) the existing state-of-the-art (SKCV, blocking, KCV); (2) R-KCV that only considers road distance and; (3) T-KCV that only considers travel time. Figure 3 provides a flow diagram of the entire experimental validation process for all spatial k-fold methods: SKCV, R-KCV, T-KCV and RT-KCV. For our case studies we assume the resulting model to be Kriging-based and the validation metric to be NRMSE.

Definition. Assume a data point $\mathbf{d}_i = (\mathbf{x}_i, y_i, \mathbf{c}_i)$ where $\mathbf{x}_i \in \mathbf{R}^D$ is a feature vector, $y_i \in \mathbf{R}$ is the response/target and $\mathbf{c}_i \in \mathbf{R}^2$ is the geographical coordinate vector of the ith data point in dataset $\mathbf{D}=\{d_1, d_2, \ldots, d_n\}$. Additionally, consider a set of distance matrices $\mathcal{M} = \{\text{Road Distance (RD), Travel Time (TT)}\}$, where the ijth position within each matrix refers to the shortest road distance or average time taken to travel by car (including legal and social constraints) between points i and j respectively. We define $\rho \in \mathbf{R}^+$ to be the dead-zone radius and $\nu = \{\nu_1, \ldots, \nu_k\}$ to be the set of KCV folds. Vector $\hat{\mathbf{y}} \in \mathbf{R}^n$ is the predicted response values from model $\mathcal{F}$ (intro-



(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 are based on accessibility.

Figure 2.: Road distance versus Euclidean dead-zone example.

duced in Section 5). Additionally, α_1 and α_2 are user defined weightings to calculate to what extent road distance and travel time form the basis of our RDTT matrix. Finally, a validation metric (in our case 'NRMSE') is selected to measure model accuracy.

Determining the Optimal Dead-Zone. RT-KCV and variants have a free parameter which is the dead-zone distance. To the best of our knowledge, no work has been undertaken to determine what this distance should be on a case-by-case basis. The dead-zone distance directly defines the amount of SAC that will be removed and hence implicitly defines the setting (interpolation-extrapolation) that a model is expected to be in and therefore how accurate the estimated generalisation performance will be. In this research we 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

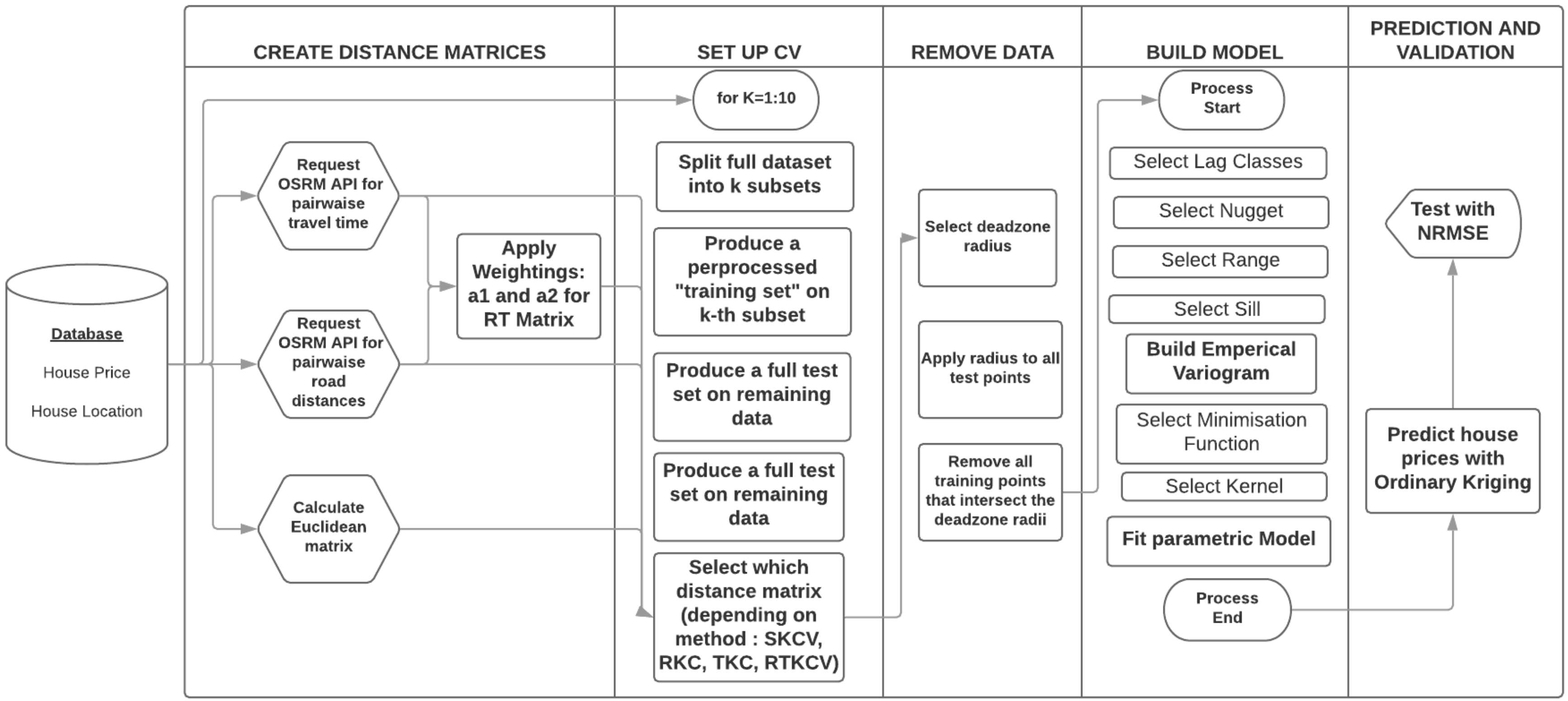


Figure 3.: Flow Diagram of SKCV, R-KCV, T-KCV and RT-KCV algorithm.

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, we select the dead-zone distance based on the heuristic presented in formula 1, which was found to perform well across all settings and datasets.

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

where μ'_{tt}, μ'_{tr}, μ'_{te} are the average distances between each point in the (1) training and test set, (2) within training set and (3) the validation test set respectively. Once the distance is selected, we 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, SKCV). We term the output of a model using this heuristic as the *mean operating point*. The purpose of this is to work out the setting that the model is training on (interpolation, extrapolation or a combination of the two).

5. Urban Case Studies

Given that (non-Euclidean) distance is shown to be the single most influential variable in urban house price predictions (Crosby *et al.* 2016, 2018), our first case study builds a valuation model with no covariates on a set of 3,413 residential sold house prices in Coventry, UK for 2016. Our second case study utilises historic traffic flow information on 711 sensor locations in Birmingham, UK.

5.1. *Related Applications*

House price predictors, previously defined as automated valuation models (AVMs), attempt to exploit data to reliably understand the value of real estate over a large area where market behaviour may differ significantly (Crosby *et al.* 2018). Most contemporary machine learning based AVMs are hedonic (McClusky and Borst 2007). For example some models utilise topography, natural geography (Kok *et al.* 2011), building footprints (Pace *et al.* 1998) and crime data (Thaler 1978). Space (Crosby *et al.* 2016) and time (Huang *et al.* 2010) also contribute to hedonic models, inferring up to 71% of the price of a residential property. As such, our paper considers spatial modelling only.

Typical spatial models attempt to obtain a description of spatial continuity within a dataset by calculating the variance of pairwise distances (Cressie 2015, Matheron 1963). One such example is a semivariogram which supports Kriging. Given our interest in road networks, Crosby *et al.* (2018) utilises OpenStreetMap to find which Minkowski P value is most related to road distance, travel time and a linear combination of the two; a linear combination of the two produces a significantly improved AVM. This infers that road networks may contain more SAC than Euclidean distance and can be used to support our hypothesis.

Traffic flow prediction has a wide range of applications, including assessing potential designs for new road layouts, reducing accident hotspots and short-term prediction of traffic congestion (Yin *et al.* 2002, Sun *et al.* 2006). Temporal traffic predictions have been generated utilising ARIMA (Williams and Hoel 2003), Markov chains (Yu *et al.* 2003), Bayesian Belief Networks (BBN) (Sun *et al.* 2006) and Artificial Neural Networks (ANN) (Duan *et al.* 2015), with mean absolute percentage errors of 8.6% (Wang *et al.* 2011). Most notably, Zou *et al.* (2012) supports our research in that they (1) identify spatial modelling as being the most appropriate for urban traffic flow prediction, (2) note that Kriging performs best for their case study and (3) show that an approximate road distance metric better infers SAC than Euclidean for their model.

5.2. *The base Kriging predictor*

5.2.1. Ordinary Kriging

For both case studies, we consider Ordinary Kriging a spatial predictor which accounts for spatial covariance based on observed pairwise distances. We use Ordinary Kriging as a simple and widely utilised spatial statistical model in order to demonstrate the benefits of RT-KCV. The method begins by calculating an empirical semivariogram based on a user defined lag, produced by the pairwise distances of all points in a dataset. Formally, let S be a set of spatial locations defined in Section 4 such that $s_i : i = 1, \ldots, n$ where $s \epsilon S \subset \mathbf{R}^d$ are known and $Z(s) : s \subset \mathbf{R}^d$ is a real valued stochastic process over random fields. To predict some value $Z(s_0)$ at location s_0 from observed values $Z(s_i) : i = 1, \ldots, n$, the data must represent a complete sampling of a single realisation (second order stationary):

$$E[Z(s)] = \mu \text{ for all } s \epsilon S \tag{2}$$

and

$$cov[Z(s_1), Z(s_2)] = C(s_1 - s_2) \text{ for all } s_1, s_2 \epsilon S \tag{3}$$

Suppose a function $\gamma(h)$ is the semivariogram and that the data has a stationary covariance ($\mathbf{C}$), then the semivariance is related to the covariance function with a nugget, sill and range (Cressie 1988). This stands such that h is the distance between points s_i, s_j. Given this, the covariance function is estimated by fitting a parametric model to the calculated semivariance, typically using least squares.

As such, Ordinary Kriging estimates value $Z(s_0)$ at point s_0 with the known variogram implicitly evaluating the mean of a moving neighbourhood (Wackernagel 1995). Generally, the local mean of a Kriging estimate is calculated, then a simple estimator is taken from a 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 linear weights λ_i, such that

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i) \tag{4}$$

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 the same constant, this guarantees uniform unbias (Equation 2). We assume the data to be part of a realisation of an intrinsic random function with $\gamma(h)$. Given that the expectation of each increment is 0, an unbias with unit sum weights must be calculated

$$\begin{aligned} E[\hat{Z}(s_0) - Z(s_0)] &= E[\sum_{i=1}^{n} \lambda_i Z(s_i) - Z(s_0) * \sum_{i=1}^{n} \lambda_i] \\ &= \sum_{i=1}^{n} \lambda_i E[z(s_i) - z(s_0)] = 0 \end{aligned} \tag{5}$$

An optimal Kriging predictor is one which minimises the mean-squared prediction error (Cressie 2015)

$$\sigma^2(s_0) = E[(\hat{Z}(s_0) - Z(s_0))^2] \tag{6}$$

over the class of linear predictors $\sum_{i=1}^{n} \lambda_i = 1$

$$2\gamma(h) = var(Z(s+h) - z(s)), \quad h \in \mathbf{R}^d \tag{7}$$

Hence, given Equation 6, one must minimise (differentiate and equate to 0) equation 8

$$E\big(Z(s_0) - \sum_{i=1}^{n} \lambda_i Z(s_i)\big)^2 - 2m\big(\sum_{i=1}^{n} \lambda_i - 1\big) \tag{8}$$

with respect to $\lambda_1, \ldots, \lambda_n$, and the Lagrange multiplier m (ensuring $\sum_{i=1}^{n} \lambda_i = 1$), which shows that optimal $\lambda_1, \ldots, \lambda_n$ can be obtained from

$$\lambda_0 = \Gamma_0^{-1} \gamma_0 \tag{9}$$

This produces an Ordinary Kriging system of:

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

The weights λ_i are assigned to the $Z(s_i)$ to show 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) where n is the total number of points. Given that $Z(s_i)$ is second order stationary and the distance matrix is positive definite, then

$$Var(\hat{Z}) = \sum_i \sum_j \lambda_i \lambda_j C(h) \geq 0. \tag{10}$$

Commonly applied covariance ($\mathbf{C}$) functions have been built based on Euclidean distances to ensure this holds.

Table 1.: A subset of restrictions to road network and travel time in OSRM calculations.

Restriction Type	**Description**
Barrier	(Rising) bollard, cattle grid, toll booth ...
Restriction	Motor vehicle, vehicle, permissive, private, forestry, emergency ...
Speed Profile	Motorway, trunk, primary, secondary, residential, unclassified ...
Tracktype Speeds	Grade 1-5, intermediate ...
Maxspeed	Urban, rural, trunk, motorway ...
U-Turns & Signals	Time in seconds
Oneway	Boolean, y/n

5.2.2. 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 OSM (see Table 1). From their API, we are able to calculate an nxn distance matrix $\mathcal{M}$ for all points. Our combination (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. In Section 6, we speculate that a future avenue of research is to build a heuristic/metric which can optimise these weightings.

5.3. *Validation*

We evaluate our proposed KCV methods against the current state-of-the-art. The primary purpose of any (cross-)validation procedure is to estimate a model's generalisation performance to unseen data. As such we propose three settings (interpolation, extrapolation and that between) over two case studies (house price and traffic flow prediction), each with a number of KCV methods (KCV, SKCV, R-KCV, T-KCV, RT-KCV and blocking KCV). In order to evaluate the validation techniques, we compare each method against a 'ground truth' test set, which are visualised in Figure 5(a)-(f), showing a set of six simulated real-world scenarios. Each of these ground truths show the results that we would get if we ran the experiment in each interpolation to extrapolation setting and hence validates how well cross-validation would estimate the generalisation performance of our model to unseen. As such, the cross-validation method which performs closest to our ground truth is the best performing. For robustness, we keep the same ground truth test in all experiments within a case study and the same validation test sets for all cross-validation approaches.

We also compare our method against the most popular competitor approach; blocking cross-validation (Roberts *et al.* 2017). Our 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 if the total number of points is odd). 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 (256 and 72 for house prices and traffic flow respectively). See Figure 4 for a visualisation of blocking on a subset of 3 folds. Each coloured block represents a different fold and the point colours refer to the different house prices in the data set where dark red is represents the highest house prices and white shows the lowest house prices. The blocks are different sizes, so that each block considers an equal number of points.

In order to account for any variability due to the choice of the ground truth test set, we resample multiple ground truth test sets for our non-interpolation settings (settings B and C; defined in Sections 5.4 and 5.5 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 we present three approaches to validate our KCV methods against the ground truth.

Model Validation: Our Kriging model is validated against the normalised root mean squared error (NRMSE) which takes the square root of the mean squared error and is then normalised by the difference of the y values

$$NRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{y_{max} - y_{min}} \tag{11}$$

where $\hat{y}_i$ is the predicted value of y_i and y_{max}, y_{min} are the maximum and minimum values of y in the dataset.

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 (the NRSME of the ground truth case) with the fewest training points removed by dead-zones. We 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 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.

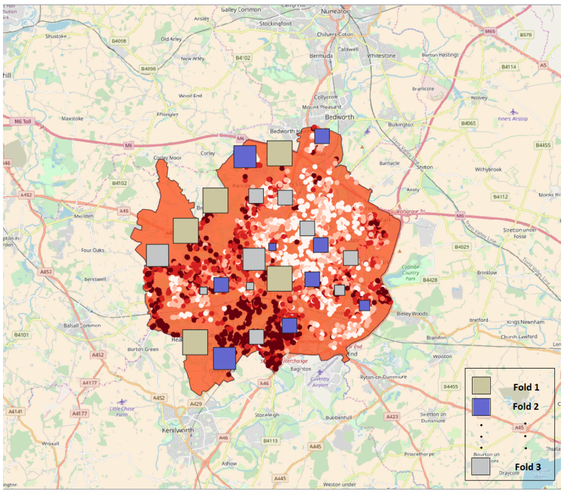
5.4. ***Case Study 1 - House Price Prediction***

Our house price data is openly sourced from Her Majesty's Land Registry (HMLR) which provides all sold residential properties in England and Wales since 1995. In addition, Ordnance Survey (OS) provides the location of each address. We select all freehold houses between $01-01-2016$ and $01-06-2017$ in the city of Coventry, accounting for 3,669 properties in total. The precise dataset used in this study matches that from Crosby *et al.* (2018), which shows a strong Moran's I result of $I_{observed} = 0.1559136 >> I_{expected} = -0.00267094$ with a P-value ~ 0. This complements our findings shown in Figure 1 and confirms that our experiment can be spatially modelled (contain global SAC) according to a standard Moran's I test (Moran 1950). These results allow us to reject the null hypothesis that there is no SAC for house prices at alpha = 0.05.

To determine our ground truth test sets, we consider 3 settings; pure extrapolation, mixed interpolation/extrapolation and pure interpolation to test our newly defined methods across a range of experiments (see Figures 5 (a)-(c)).

Setting A - Pure Extrapolation: We train on all data that sit within the Office of

Figure 4.: Blocking Cross Fold validation with equal test sets.



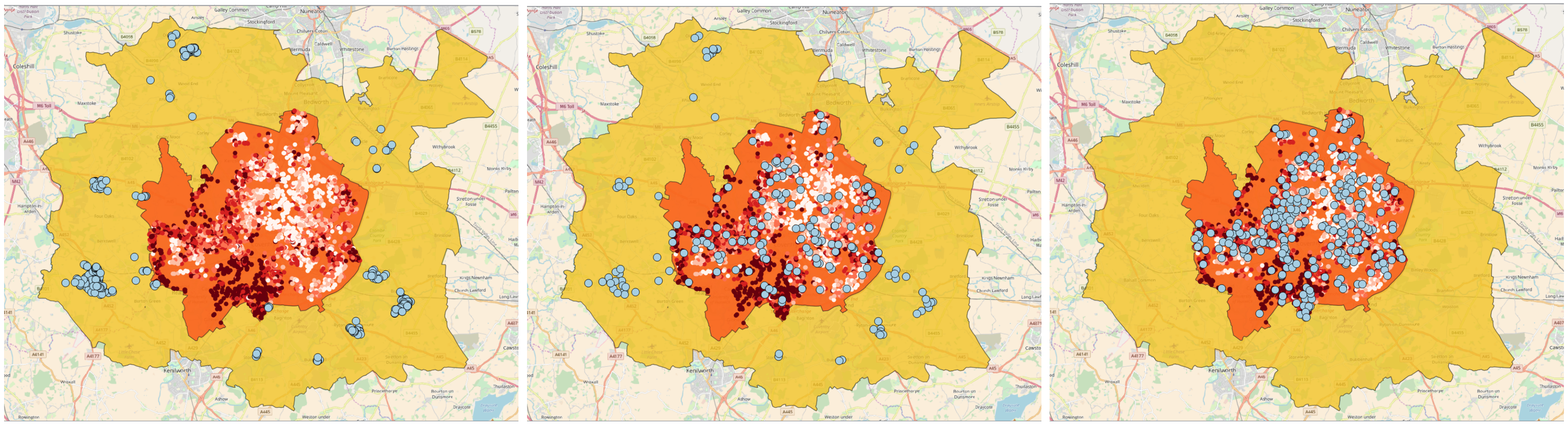
National Statistics (ONS) classified Built Up Area (BUA), accounting for 3,413 houses. The remainder are removed for testing in our ground truth test sets which account for 256 points. To simulate extrapolation fully, we confirm that our train and hold out sets are not correlated (i.e. $SAC_{traintest} \sim 0$). Again, a standard Moran's I test is conducted between both datasets showing a weak spatial relationship such that $I_{observed} = 0.020206$ and $I_{expected} = 0.019014$. As such, we confirm that our 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: We train on data that sit within the Coventry BUA only. For our ground truth scenario, half the test set sits within the BUA and half sits outside, thus the training set consists of 3,291 houses.

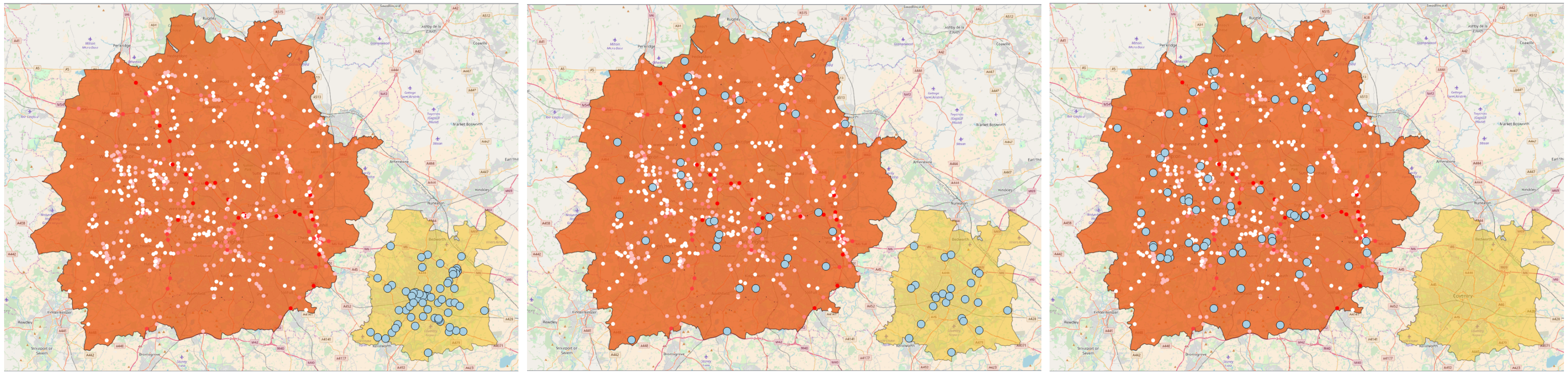
Setting C - Interpolation: We train on data that sit within the Coventry BUA. For our ground truth scenario, all the test points lay within the BUA, accounting for a training set of 3,163 houses.

5.4.1. 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(a)-(c) show the NRMSE value for each cross-validation method (KCV, SKCV, 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, showing that RT-KCV consistently outperforms all other approaches in all settings (that is it approaches the ground truth with fewer points removed). Each vertical bar represents the spread of results across all experiments. Notably, RT-KCV requires only 8 points to be removed to ensure the same SAC removal as 201 points for SKCV in our interpolation setting. In addition, Table 2 shows that RT-KCV consistently generalises 50%, 80% and 100% of the ground truth with fewer points removed than any other method. Finally, our 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, we determine the difference in the estimated NRMSE values (0.11162 ,0.128 and 0.104)

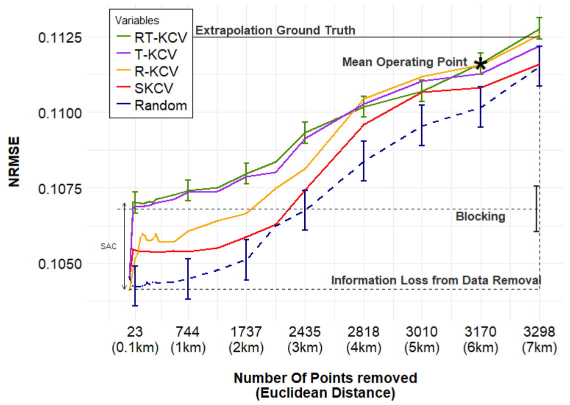


(a) House prices in Coventry - Extrapolation (b) House prices in Coventry - Mixed (c) House prices in Coventry - Interpolation

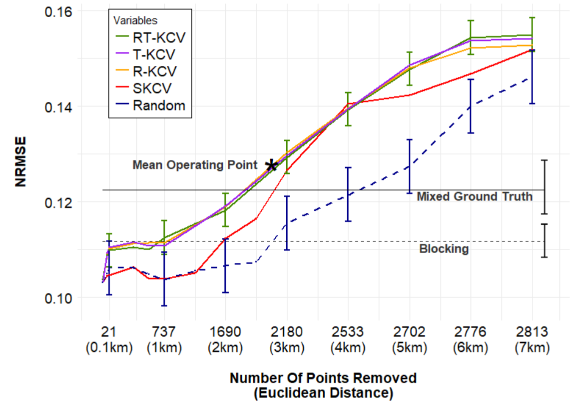


(d) Traffic flow in the West Midlands - Extrapolation (e) Traffic flow in the West Midlands - Mixed (f) Traffic flow in the West Midlands - Interpolation

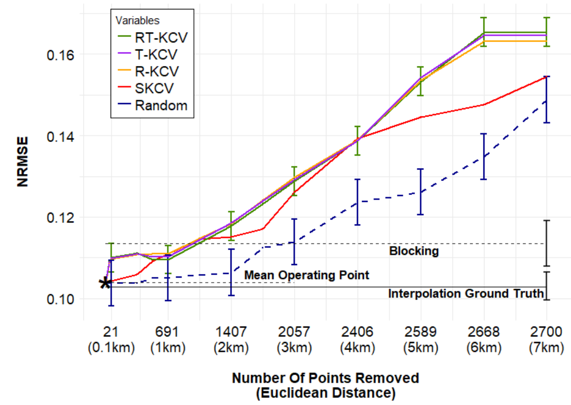
Figure 5.: Producing a ground truth training and test set: The orange space represents the training area; The yellow space represents the ground truth test area; The blue points are ground truth testing locations and, the white to red points represent the training set, where white points are the cheaper houses and lower traffic flows and the red points are the more expensive houses and higher traffic flows.



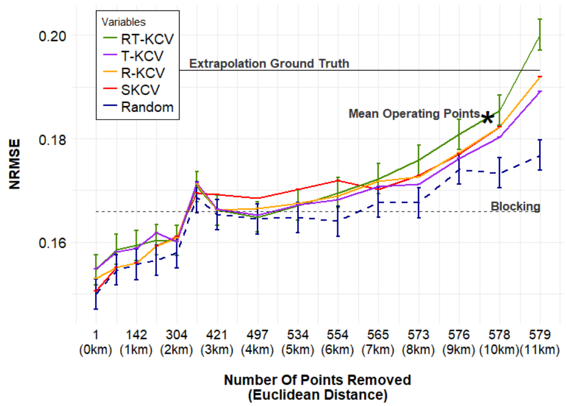
(a) Coventry house prices with all KCV methods - extrapolation.



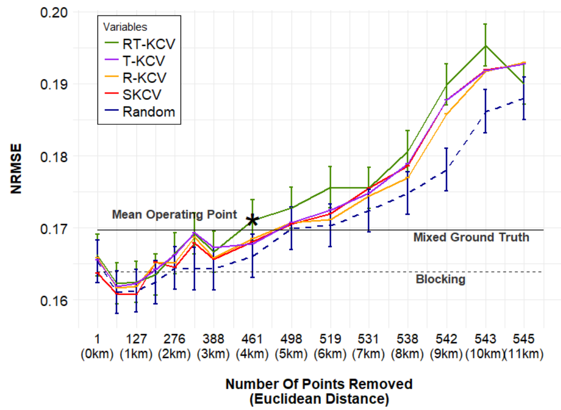
(b) Coventry house prices with all KCV methods - mixed.



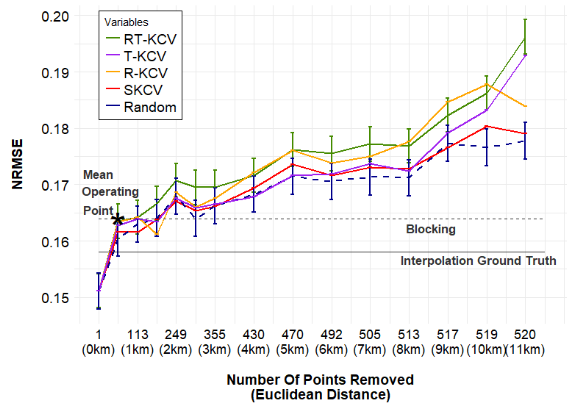
(c) Coventry house prices with all KCV methods - interpolation.



(d) Traffic flow with all KCV methods - extrapolation.



(e) Traffic flow with all KCV methods - mixed.



(f) Traffic flow with all KCV methods - interpolation.

Figure 6.: Results for both case studies. Dead-zone versus NRMSE for all KCV methods, ground truth and blocking.

Table 2.: Results : the number of points removed to reach a specific % of NRMSE for each KCV technique.

Results Table											
	Real Estate Case Study					Traffic Flow Case Study					
	Random	Previous Work (Pohjankukka *et al.* 2017)	Our Work			Random	Previous Work (Pohjankukka *et al.* 2017)	Our Work			
	KCV	SKCV	R-KCV	T-KCV	RT-KCV	KCV	SKCV	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**	

and the ground truth values (0.1125, 0.1225974 and 0.1135) are relatively small compared to SKCV and blocking. A t-test, with a t-value of 0.01, shows that for all settings, the number of points that are removed from the training set are significantly less with our new RT-KCV approach compared with the previous state-of-the-art. Table 2 shows in bold the best performing cross-validation approaches for each scenario.

5.5. *Case Study 2 - Traffic Flow Prediction*

Our predictor considers the total average daily traffic flow between $01 - 01 - 2016$ and $01 - 06 - 2017$ for Birmingham, UK accounting for 711 sensors as supplied by the ‘Highways Agency England’. The hold out test set (i.e., ground truth) considers Coventry, UK with 63 sensors in the same time period.

To confirm SAC we conduct a standard Moran’s I test where $I_{observed} = 0.1604474 >> I_{expected} = 0.0002727025$ with a P-value of 0. These results also allow us to reject the null hypothesis that there is no SAC present at alpha = 0.05. To determine our ground truth test sets, we consider 3 settings; pure extrapolation, mixed interpolation/extrapolation and pure interpolation to test our newly defined methods across a range of experiments (see Figures 5 (d)-(f)).

Setting A - Extrapolation: We 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, we confirm that our training and hold out sets are not correlated (i.e. $SAC_{traintest} \sim 0$). A standard Moran’s I test is conducted between both datasets showing a weak spatial relation such that $I_{observed} = -0.008960041$ and $I_{expected} = 0.000201$. As such, we confirm that our method can be tested against the split data for extrapolation generalisability, see Figure 5(d) for a visual representation.

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

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

A Competitor Case for Comparison - Blocking: Our 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 if the total number of points is odd) and also sums to the same sized test set as all other experiments (72 points). We only apply this in settings B and C because setting A contains no test points within the training set.

5.5.1. 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(d)-(f) show the NRMSE value for each cross-validation method (KCV, SKCV, 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 - this does not refer to it’s ability to present the optimal result, but instead the most realistic, compared to our earlier defined ground truth. Notably, the benefits of RT-KCV to this case study compared with the house price case study is less significant. This can be explained by the weaker spatial correlation as seen by our semivariogram in Figure 1 and by our Moran's I value. Finally, our dead-zone radius heuristic estimates that 577, 458 and 87 points need to be removed for extrapolation, mixed and interpolation respectively. Once implemented, we 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 SKCV and blocking (with the exception of interpolation which is negligible). A t-test, with a t-value of 0.01, 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 our new RT-KCV approach compared with the previous state-of-the-art. In addition, Figure 5 empirically demonstrates a significant estimation of generalisation improvement, because we see that the 'mean operating point' (our newly defined measure of generalisation performance) is the same or closer to the ground truth in all scenario's of extrapolation to interpolation, as compared with SKCV and blocking (the current state-of-the-art) for both case studies.

6. Conclusions

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 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 we demonstrate that our 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 in KCV approaches (SKCV). We 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, we also show that RT-KCV is a superior approach for cross-validation and estimating model generalisation than blocking cross-validation, the current state-of-the-art.

We recommend that RT-KCV be considered 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 accounts for such correlations (for example in Kriging). However, in some cases, Euclidean (or other) distances may be more appropriate, such as the migration of birds or direction of air pollution. We note that standard KCV is only appropriate for pure interpolation where the internal dependence structure would otherwise be present in the locations with unknown values. Notably, we show that, for urban data, a combination of road distance and travel time capture SAC better than Euclidean distances.

Further avenues for research include: (i) Developing techniques to better map SAC in other dependent datasets, such as 'stream' distances (along a river or canal) or coastal distances. These methods may be more suited to ecological problems, for example the migration patterns of animals; (ii) Optimising the operating point on the RT-KCV curve to better match the ground truth performance. This would allow researchers to more accurately determine the effect of space on any cross-validation technique; (iii) Learning the convex combination parameters for the combined RDTT distance so that we are better able to measure a person's perception of distance in a cityscape and; (iv) Considering the effect of alternative mode's of transport in certain cities, for example trains, buses, walking and bicycles, especially those cities with traffic-free zones and a heavy reliance on bikes and public transport.

Acknowledgements

We would like to thank the Engineering and Physical Sciences Research Council (EPSRC) Centre for Doctoral Training in Urban Science (EP/L016400/1) and the Alan Turing Institute (EP/N510129/1) grants. In addition, we are supported by the Lloyds Register Foundation programme on Data Centric Engineering. Finally, our gratitude goes to all the open source mapping contributors.

References

Anselin, L., 1995. Local indicators of spatial association (LISA). *Geographical analysis*, 27 (2), 93–115.

Christaller, W., 1933. *Die Zentralen Orte in Suddeutschland*. Jena: Gustave Fischer.

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

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

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

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

Cressie, N., 2015. Statistics for spatial data. *Terra Nova*, 4 (5), 613–617.

Crosby, H., *et al.*, 2018. Road Distance and Travel Time for an Improved House Price Kriging Predictor. *Geo-spatial Information Science*, 21 (03), 185–194.

Crosby, H., *et al.*, 2016. 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.

Duan, Y., *et al.*, 2015. Traffic Flow Prediction With Big Data: A Deep Learning Approach. *IEEE Transactions on Intelligent Transportation Systems*, 16 (2), 865–873.

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

Fotheringham, S. and Rogerson, P., 2013. *Spatial analysis and GIS*. CRC Press.

Getis, A., 1995. Cliff, A.D. and Ord, J.K. 1973: Spatial autocorrelation. London: Pion. *Progress in Human Geography*, 19 (2), 245–249.

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

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

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

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

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

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

Kohavi, R., *et al.*, 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *In*: *Ijcai*, Vol. 14, 1137–1145.

Kok, *et al.*, 2011. Economic geography, jobs, and regulations: the value of land and housing. *AREUEA Meetings Denver*.

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

Le Rest, K., *et al.*, 2014. Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Global ecology and biogeography*, 23 (7), 811–820.

Matheron, G., 1963. " Principles of geostatistics", Economic Geology, pp 1246-1266. .

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

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

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

Pace, R.K., *et al.*, 1998. Spatiotemporal autoregressive models of neighborhood effects. *The Journal of Real Estate Finance and Economics*, 17 (1), 15–33.

Pohjankukka, J., *et al.*, 2017. Estimating the prediction performance of spatial models via spatial k-fold cross validation. *International Journal of Geographical Information Science*, 31 (10), 2001–2019.

Prudhomme, M., 2018. Press Release of the World Urbanization Prospects 2018. *https://population.un.org/wup/Publications/*.

Roberts, D.R., *et al.*, 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40 (8), 913–929.

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

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

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

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

Wackernagel, H., 1995. *In*: *Ordinary Kriging.*, 74–81 Berlin, Heidelberg: Springer Berlin Heidelberg.

Wang, J., Shang, P., and Zhao, X., 2011. A New Traffic Speed Forecasting Method Based on Bi-Pattern Recognition. *Fluctuation and Noise Letters*, 10 (01), 59–75.

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

Yin, H., *et al.*, 2002. Urban traffic flow prediction using a fuzzy-neural approach. *Transportation Research Part C: Emerging Technologies*, 10 (2), 85 – 98.

Yu, G., *et al.*, 2003. Short-term traffic flow forecasting based on Markov chain model. *In*: *Intelligent Vehicles Symposium, 2003. Proceedings. IEEE*, 208–212.

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

Zou, H., *et al.*, 2012. 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.