A Spatio-Temporal, Gaussian Process Regression, Real-Estate Price Predictor

Henry Crosby
Department of Computer Science
University of Warwick
Coventry, UK

Paul Davis
Assured Property Group
Unit 46 Innovation Centre
Warwick Technology Park
Warwick, UK

Theo Damoulas
Department of Computer Science
& Department of Statistics
University of Warwick
Coventry, UK

Stephen A. Jarvis
Department of Computer Science
University of Warwick
Coventry, UK

ABSTRACT

This paper introduces a novel four-stage methodology for real-estate valuation. This research shows that space, property, economic, neighbourhood and time features are all contributing factors in producing a house price predictor in which validation shows a 96.6% accuracy on Gaussian Process Regression beating regression-kriging, random forests and an M5P-decision-tree. The output is integrated into a commercial real estate decision engine.

Keywords

Gaussian Process Regression, Universal Kriging, Machine Learning, Space Time Cube, Real Estate Valuation

1. INTRODUCTION

The primary aim of real-estate value prediction differs significantly from one stakeholder to another: (1) A developer is looking simply to maximise their returns by purchasing undervalued land or property; (2) Lenders and economists seek to minimise their risk and hence are interested primarily in the market form efficiency of real estate; (3) Home buyers have non-return related priorities, such as life-style suitability and location. Varying motives, macro-environmental factors and data sparsity are just some of the reasons why real-estate valuation remains a challenging task. Additionally, unprecedented leverage (for example 100% loans) and market potential (a 4.5% increase in residential property prices in the UK in 2015) [8] are some of the reasons why real-estate value prediction can be so valuable. In a recent study [4] it was shown that the most accepted UK real-estate prediction models have been designed by the Office for National Statistics (ONS), the Land Registry, the Halifax and Nationwide Building Societies, and the LSL Academy (formed by Pink and Aviva), which all utilise the Land Registry' open source sales data and/or private mortgage archives. The most popular valuation models include (1) median sales prices (2) repeat sales and (3) hedonic regressions. Inconsistent outputs between these models provide uncertain accuracy of these approaches [4]. This research, in partnership with the Assured Property Group, aims to provide a real-estate valuation model that delivers accurate and consistent valuations for the whole of England and Wales. The benefit of this work over others is its ability to predict individual house prices based on space, time, economic and neighbourhood features, rather than previously discussed spatial aggregates with a comparably uninformed feature set.

The remainder of the paper is as follows: Section 2 reviews the most successful residential value predictors published to date; Section 3 describes the scientific methodology; Section 4 details the results of our validations; Section 5 concludes the paper and presents avenues for further research.

2. BACKGROUND RESEARCH

From the perspective of lagged returns, the most popular valuation models employ constant-quality indices, notably repeat sale regression (RSR) [1]. As with most regressions the RSR predicts through the application of cross-sectional regression prior to applying ordinary least squares. In a comparison study of methodologies, the RSR algorithm was tested on 53,000 properties in and around Chicago and only gained an R^2 of 0.34 [7]. The key reason for such a poor fit was due to the assumption that property characteristics have no bearing on its price.

[9] introduced a RIPPER regression on 5,359 houses in Fairfax county. Similar to our research, this paper introduced

SIGSPATIAL'16, October 31-November 03, 2016, Burlingame, CA, USA

DOI: http://dx.doi.org/10.1145/2996913.2996960

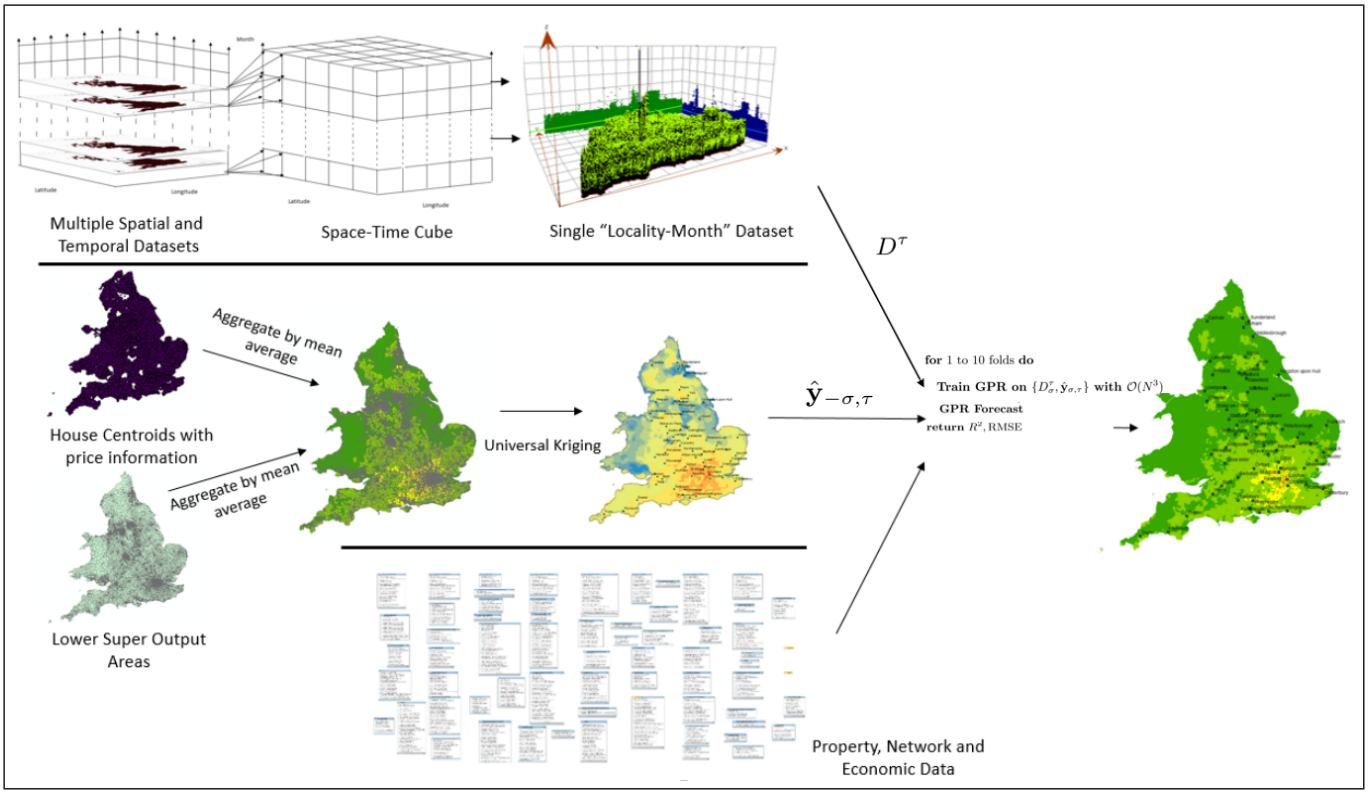

Figure 1: Process Diagram Corresponding to the Space, Time, Property, Network, Time Algorithm Detailed in Algorithm 1.

a number of property characteristics: size of building footprint, presence of a garage, the number of parking spaces, and the size of the entire title. Additionally, they introduced a number of environmental factors such as local school performance, mortgage contract rate, list month, list price, year built and number of days on the market. Their approach produced a relative error of 0.248. The results of this paper are informative; however, the size and variation of the data is relatively limited compared to that available via open source in the UK. Hence, our research includes variables such as listed building status, council tax band, supermarket distance, flood risk and coastal proximity across a larger space.

With regard to the prediction of spatially dependant datasets, spatial statistics and the removal of the IID assumption, the most relevant research in this area is that of [6]. Their research, based on 30,000 properties over six years in Lucas County, showed that SDM-MISS (an extension to the Spatial Durbin Model) removed 75% of the error between least square prediction errors and those from the popularly employed spatial autoregressive model. Our research utilises similar Gaussian process approaches, with the benefit of including network, economic and property features.

From the perspective of modelling data with (spatial) dependencies, [5] attempted a study between four kriging techniques: detrended kriging, universal kriging (UnK), detrended co-kriging and universal co-kriging on 1,707 households in Austria, yielding an R^2 of 0.66 on the best model (universal co-kriging). Our paper includes spatial predictions as a single feature in the final Regression (GPR).

3. SCIENTIFIC METHOD

Our method introduces a novel, four-stage, methodology for real-estate valuation as seen in figure 1:

Stage 1 (Temporal Interpolation): As defined by the Office of National Statistics (ONS) we tested LSOA's, OA's and postcode areas to identify the granularity of space required to enable an optimally representative space-time interpolation. The interpolation provided a new dataset D^τ where each point is time singular in price. The mean value of each area was calculated and then extended on each property in the land registry's sales dataset. All properties sold in April and May 2016 were used as the test set, this amounted to over 12,000 transactions. Output Area's (OA's) were utilised, yielding an R^2 value of 0.71.

Stage 2 (Spatial Dependency Identification): This stage overcomes the identical and independent distribution assumption put forward in most statistical techniques. Common spatial interpolation techniques are based on Gaussian processes set by some prior covariance function, known as Kriging. In our case, Universal Kriging (UnK) was utilised on a stratified sample of D^τ named D^τ_σ and then interpolated to produce ouput $\hat{y}_{-\sigma,\tau}$. UnK was used because unlike the popularly employed method of Ordinary Kriging it assumes non stationarity. Uniquely, UnK considers the spatial correlation between the points that need to be interpolated and their neighbouring points [11]. Four covariance functions (kernels) were tested with the UnK algorithm; Epanechnikov, Gaussian, Polynomial and Exponential. The best performing UnK method employed a fifth order polynomial

kernel function. Finally Inverse Distance Weighting (IDW) and Emperical Bayesian Kriging (EBK) were tested against the UnK, see Table 1 for a comparison of results.

Result	IDW	EBK	UnK
RMSE	104383.2	94810.6	94443.1
R^2	0.816	0.836	0.839

Table 1: Spatial Statistic Model Performance Comparison

Stage 3 (property, network and economic features): A pre-defined set of training features were agreed in co-operation with industry experts alongside the removal of all features with a Pearson correlation coefficient $\geq 85\%$ of each other. Property features, such as building footprints, building height, title size, property type (detached, terraced, apartment, etc.), freehold status and old/new-build status were utilised. Network features included proximity to schools and the performance of the closest primary and secondary school, proximity and usage of the closest train station, traffic flow passing the property, the population density within (1) the postcode and (2) within 250, 500 and 1,000 meters of the property. Finally, economic features such as variable mortgage interest rate, total number of houses sold each month, inflation and USD exchange rate were included.

Stage 4 (Gaussian Process Regression): We trained a model to predict house prices based on the dataset $\{D^{\tau}_{\sigma}, \hat{\mathbf{y}}_{\sigma,\tau}\}$ with a GPR, such that a model $f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k_{\text{GPR}}(\mathbf{x}, \mathbf{x}'))$ is built to provide a prediction of all house prices based on space, time, property, economic and netowrk features. A Gaussian Process (GP) is a powerful non-parametric Bayesian model, specified by a mean and a covariance (kernel) function. 16,000 location-stratified instances were trained on the GPR due to its computational complexity of $O(N^3)$ in the number of points [10]. The covariance function (kernel) chosen was a Radial Basis (Gaussian) function. The reason for including this extra step, rather than undertaking Kriging with an external drift, was to provide kernel flexibility which otherwise might not have been applicable; for example it will be seen that a separate kernel was utilised in the following GPR to that of the kernel in our UnK in Stage 2. All four stages are described using pseudo code in Algorithm 1.

4. RESULTS

Table 2 shows the results for each stage in the analytic pipeline. The columns titled 'Stage 3' and 'Stage 4' include a comparison of results on different machine learning regressions, showing that the GPR outperformed Regression-Kriging (R-K), a Random Forest (RF) and an M5P decision tree on both stages for both the R^2 and $RMSE$ validation metrics. In the final two stages, a ten-fold stratified sampling technique was implemented and the average result for each fold was calculated; the standard deviation between each fold was 4485.104. Figure 2 visualises the GPR's prediction versus actual price for all properties trained and tested. The models t-value and p-value were reported to be 27.9178 and $\leq 2.2e-16$ respectively, showing the statistical significance of the GPR model on the house price data. Figure 3 shows the uncertainty bounds between folds for properties between £0-£250,000 (the most dense section of the price distribution).

Algorithm 1: Space-Property-Economic-Network-Time (SPENT).

Require: $k_{\text{krig}}, k_{\text{GPR}}, \Theta_0 = \{\sigma^2_{f0}, \sigma^2_{n0}, \lambda_0\}$
Input: $D = \{\mathbf{X}^s_t, \mathbf{y}^s_t, \}_{\substack{s=\{1:S\} \\ t=\{t_0:\Delta t:T\}}}$
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
$D^{\tau}_{\sigma} \sim \sigma_{\text{stratified}}(D^{\tau})$
Spatial interpolation on held out locs
$\hat{\mathbf{y}}_{-\sigma,\tau} \leftarrow UnK(k_{\text{krig}}, \mathbf{y}_{\sigma,\tau})$
for 1 to 10 folds **do**
    Train GPR on $\{D^{\tau}_{\sigma}, \hat{\mathbf{y}}_{\sigma,\tau}\}$ **with** $\mathcal{O}(N^3)$
    $f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k_{\text{GPR}}(\mathbf{x}, \mathbf{x}'))$
    $\lambda, \sigma^2_f, \sigma^2_n \leftarrow \underset{\lambda,\sigma^2_f,\sigma^2_n}{\text{argmax}} \ \log p(\mathbf{y}|\mathbf{X})$ Type-II ML
    GPR Forecast
    $p(\mathbf{f}_*|\mathbf{X}_*, \mathbf{X}, \mathbf{f}) \sim \mathcal{N}(\mathbf{f}_*|\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*)$
    where $\mathbf{X}_* = [\mathbf{X}_{-\sigma}, \hat{\mathbf{y}}_{-\sigma,\tau}]$
return R^2, RMSE
Finish

4.1 Lower Bound on Performance

The previously discussed GPR was trained on 16,000 instances of historic data which were mapped to the present month (May 2016) and then validated ten-fold. 32,387 of those points were sales from April and May 2016, and hence have potential to bias the output of the predictor. A second experiment was undertaken to provide a pedantically pessimistic scenario whereby the remaining 198,613 historic data points were utilised in building the training subset and the resulting algorithm was then tested on the 32,387 non-simulated data points. The R^2 and RMSE of this experiment were 0.920 and 85,021.9 respectively. In addition, a paired t-test provided a t-value of 30.2196 and a p-value of $\leq 2.2e-16$, showing a statistical significance of the GPR model against the house price data. The mean of differences was 14094.8, with a 95% confidence interval of [13180.61,15008.99]. These results show a slight decrease in accuracy, however is relatively good considering.

4.2 Interface Implementation

'NimbusMaps', developed by Assured Property Group embeds the techniques outlined in this paper. The interface, powered by GoogleMaps, provides polygons representing all the available title numbers in the UK. A customer is able to search by postcode, current location or title number. Once this title number has been selected, information including ownership details, site size, number of buildings in the title, flood risk, estimated residential value and traffic flow are returned. One application of this tool is to consider a company that develops residential estates. Assuming that the company is interested in searching for their next development opportunity, they are able to request real-estate titles within the radius of specific location, with a specified price range, with a minimum surrounding population of x and a maximum passing traffic flow of y.

5. CONCLUSIONS

In this research we: (i) Converted several discretized, non-uniform, spatiotemporal sales data-points into a single space-

	STAGE 1	STAGE 2	STAGE 3			STAGE 4			
Result	**Interpolation**	**UnK**	**RF**	**M5P**	**GPR**	**R-K**	**RF**	**M5P**	**GPR**
Sample Size	2.1m	231,000	231,000	231,000	16,000	231,000	231,000	231,000	16,000
R^2	0.710	0.839	0.871	0.911	0.902	0.831	0.906	0.967	0.966
$RMSE$	179325.8	94,443	142,916	98,609	104,256	96,029	124,412	47,527	38,011

Table 2: Chronological Performance Comparison at Each Stage of the SPENT Algorithm.

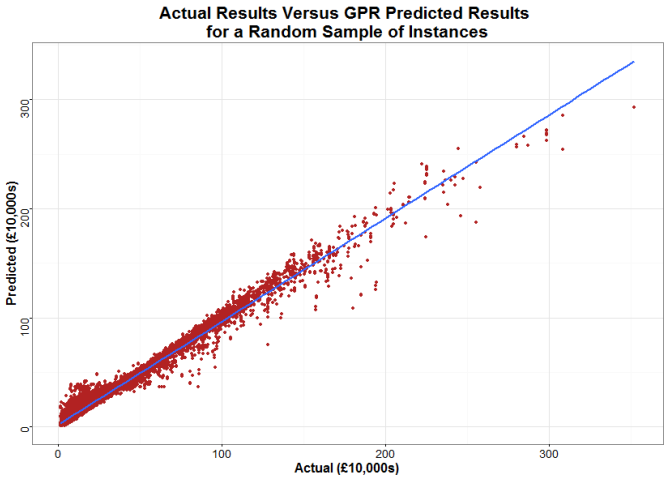

Figure 2: Actual Versus Predicted Results for Final GPR.

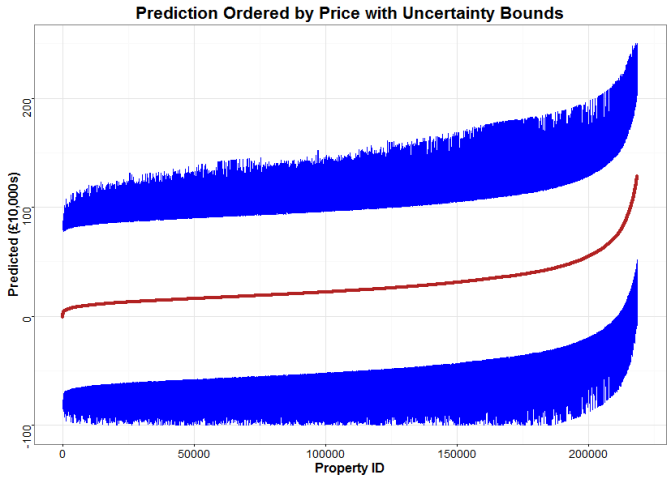

Figure 3: Prediction Uncertainty Bounds for GPR Prediction.

time cube to provide real-estate sales temporal singularity; (ii) Produced a spatially aware UnK calculation identifying house price spatial dependencies; (iii) Introduced a number of property, economic and network features to better inform the final model; (iv) Implemented each of the three previous stages into a single GPR producing an unprecedented 96.6% accuracy for real-estate price prediction (compared with previous research yielding 82% accuracy [2][3]). (v) We then implemented the results into a user-defined decision engine. Future work is to include (1) an extension to properties with no sales data, (2) seasonality inclusion and (3) scaling the GPR to allow for a larger training set.

6. ACKNOWLEDGEMENTS

We would like to thank the Engineering and Physical Sciences Research Council (EPSRC) Centre for Doctoral Training in Urban Science (EP/ L016400/ 1). T. Damoulas and S. A. Jarvis are members of the Alan Turing Institute in London, the UK's new national centre for data science.

7. REFERENCES

[1] M. Bailey, R. Muth, and H. Nourse. A regression method for real estate price index construction. *Journal of the American Statistical Association*, 1963.

[2] 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.

[3] 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.

[4] D. Chandler and R. Disney. Housing market trends and recent policies. *Institute for Fiscal Studies, London*, 2014.

[5] M. Kuntz and M. Helbich. Geostatistical mapping of real estate prices: an empirical comparison of kriging and co-kriging. *International Journal of Geographical Information Science 28:9, 1904-1921.*

[6] J. P. LeSage and R. K. Pace. Models for spatially dependent missing data. *The Journal of Real Estate Finance and Economics*, 29(2):233–254, 2004.

[7] D. P. McMillen. The return of centralization to Chicago: using repeat sales to identify changes in house price distance gradients. *Regional Science and Urban Economics*, 33(3):287 – 304, 2003.

[8] H. Meyer and H. Stewart. UK house prices: 4.5% rise in 2015 sparks policy intervention. 2015 *https://www.theguardian.com/business/2015/dec/30/uk-house-price-rise-2015.*

[9] B. Park and J. K. Bae. Using machine learning algorithms for housing price prediction: The case of Fairfax county, Virginia housing data. *Expert Systems with Applications*, 42(6):2928 – 2934, 2015.

[10] M. Seeger. Gaussian processes for machine learning. *International journal of neural systems*, 14(02), 2004.

[11] D. Zimmerman, C. Pavlik, A. Ruggles, and M. P. Armstrong. An experimental comparison of ordinary and universal kriging and inverse distance weighting. *Mathematical Geology*, 31(4):375–390, 1999.