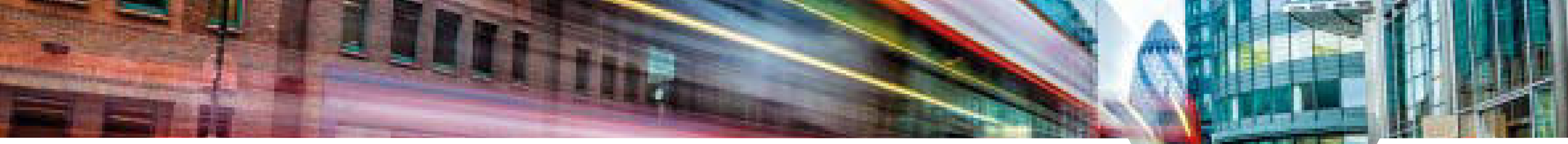


WARWICK INSTITUTE FOR THE SCIENCE OF CITIES

WARWICK

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

This paper introduces a novel four-stage methodology for real-estate valuation. The spatio-temporal Gaussian process regression is trained on a sample of 16,000 estate transactions and is validated against regression-kriging, random forests and an M5P-decision-tree with 231,000 instances utilising R^2 and RMSE. The trained model is integrated into a real estate decision engine for commercial use. Model validation demonstrates a 96.6% accuracy.

Motivations and Challenges

The primary aim of real-estate value prediction differs significantly from one stakeholder to another: (1) A developer is looking to maximise their returns; (2) Lenders 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 realestate 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) are some of the reasons why real-estate value prediction can be so valuable.

Background Reading

From the perspective of lagged returns, the most popular valuation models employ constant-quality indices, notably repeat sale regression (RSR) [1]. 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 [2]. [3] introduced a RIPPER regression on 5,359 houses in Fairfax county. Similar to our research, this paper introduced 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 limited. [4] introduced a study on 30,000 properties over six years in Lucas County, in which they 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.

Methodology

Our method introduces a novel, four-stage, methodology for real-estate valuation (see figure 1):

	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

Stage 1 (Temporal Interpolation): A space time interpolation was put forward to provide a time singular dataset (Dτ). The mean value of each area was calculated and then extended on each property in the land registry's sales dataset. The interpolation was tested on yielding an R^2 value of 0.71.

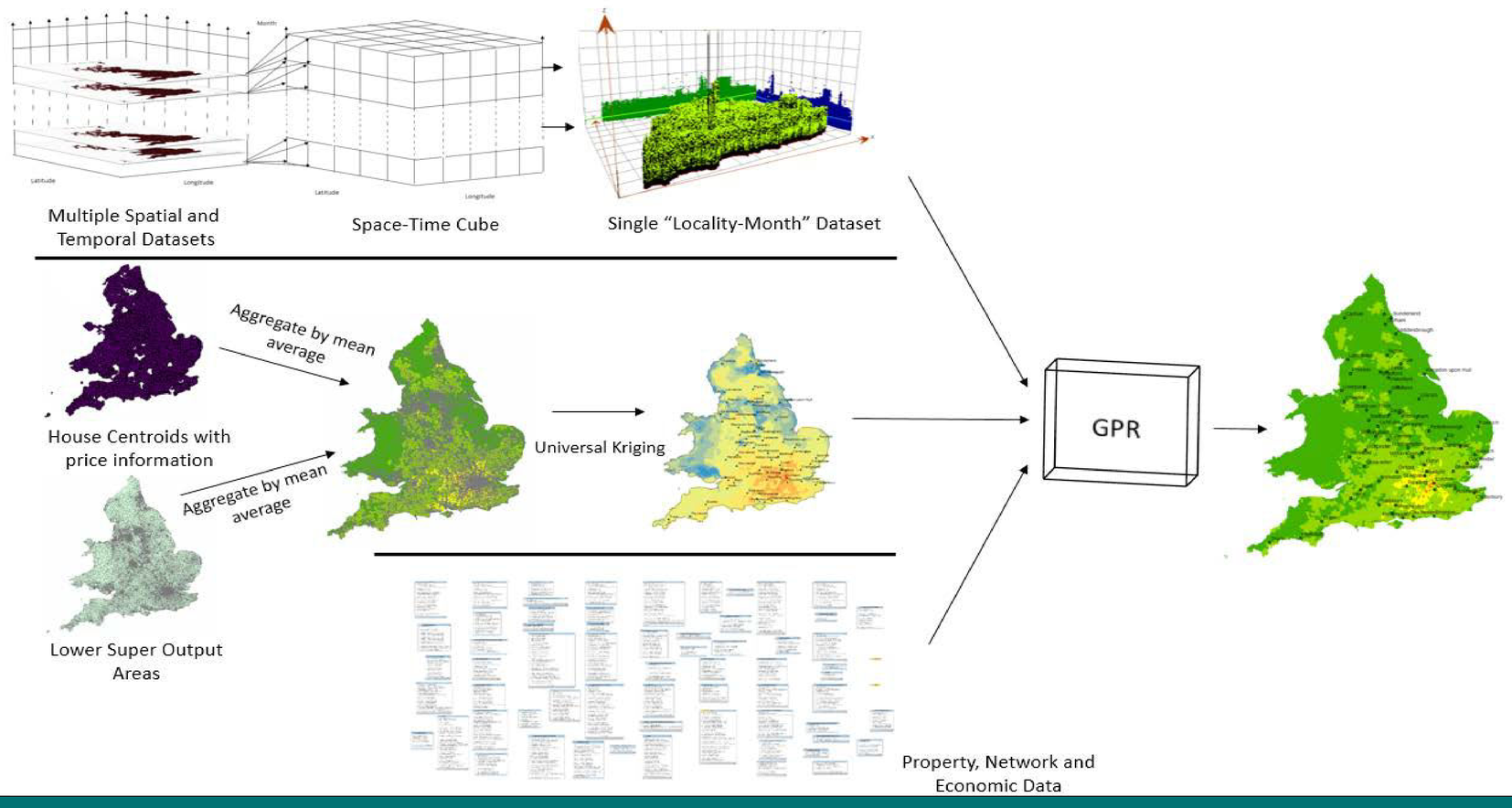
Stage 2 (Spatial Dependency Identification): Universal kriging, an interpoltion based on Gaussian processes set by some prior covariance function was utilised, this method uniquely assumes non stationarity. UnK considers the spatial correlation between the points that need to be interpolated and their neighbouring points [12]. Four covariance functions (kernels) were tested; Epanechnikov, Gaussian, Polynomial and Exponential. The best performing employed a fifth order polynomial kernel function with an R^2 of 0.839.

Stage 3 (property, network and economic features): Manual feature selection was undertaken. The remaining features include building size and height, title size, property type (detached, terraced, apartment, etc.), freehold status and old/new-build status, proximity to schools and train stations, traffic flow, population density, variable mortgage interest rates, total number of houses sold each month, inflation and GBP-USD exchange rate.

"There was a 4.5% increase in residential property prices in the UK in 2015

[6] H. Meyer and H. Stewart. 2015. UK house prices: 4.5% rise in 2015 sparks policy intervention.

Stage 4 (Gaussian Process Regression): All of the previous findings were merged into a single dataset and then a GPR was trained. 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. The covariance function (kernel) chosen was a"Radial Basis"(Gaussian) function.



Results

For comparison, two performance metrics were applied at each stage: Coefficient of Determination (R^2) and Root Mean Squared Error (RMSE) providing an absolute and relative measure for validation. Table 1 shows the results for each stage in the analytic pipeline. Figure 2 shows the actual versus predicted values of a small set of samples. 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 on. The models t-value and p-value were reported to be 27.9178 and $\leq$ 2.2e−16 respectively,

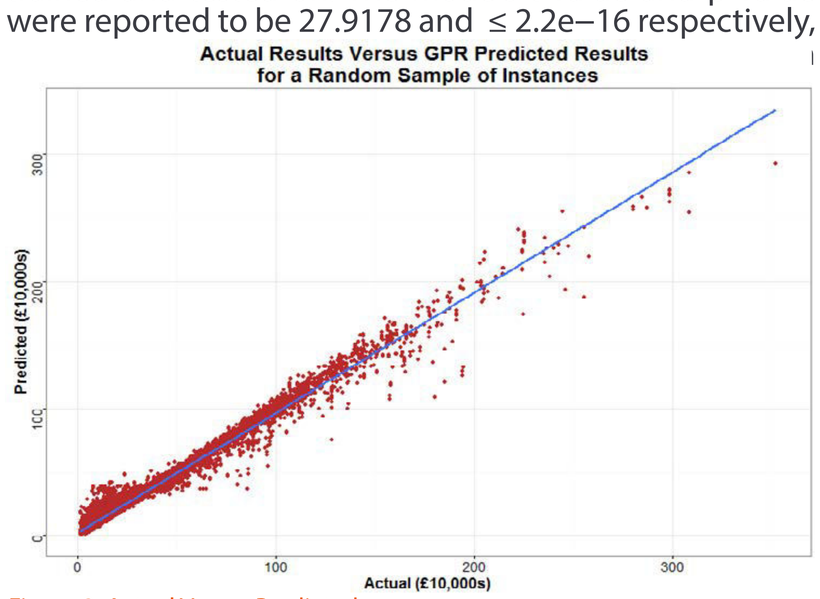


Figure 2: Actual Versus Predicted

Impact

'NimbusMaps' embeds the techniques outlined in this paper. The interface is powered by GoogleMaps with polygon overlays. A customer is able to search by postcode, current location or title number. A title number selection produces information on ownership, site size, number of buildings, flood risk, estimated residential value and traffic flow are returned.

References

[1] M. Bailey, R. Muth, and H. Nourse. A regression method for real estate price index construction. Journal of the American Statistical Association. 58 933-942, 1963.

[2] 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.

[3] 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.

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

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

[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.

Warwick Institute for the Science of Cities
The University of Warwick, CV4 7AL

wisc.warwick.ac.uk

EPSRC
Pioneering research and skills