Spatially-Intensive Decision Tree Prediction of Traffic Flow across the entire UK Road Network

Henry Crosby
Warwick Institute for the Science of Cities,
University of Warwick,
Coventry, UK

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

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

***Abstract*—This paper introduces a novel approach to predicting UK-wide daily traffic counts on all roads in England and Wales, irrespective of sensor data availability. A key finding of this research is that many roads in a network may have no local connection, but may still share some common law, and this fact can be exploited to improve simulation. In this paper we show that: (1) Traffic counts are a function of dependant spatial, temporal and neighbourhood variables; (2) Large open-source data, such as school location and public transport hubs can, with appropriate GIS and machine learning, assist the prediction of traffic counts; (3) Real-time simulation can be scaled-up to large networks with the aid of machine learning and, (4) Such techniques can be employed in real-world tools. Validation of the proposed approach demonstrates an 88.2% prediction accuracy on traffic counts across the UK.**

I. Introduction

A major concern of traffic flow prediction is to provide up-to-date, fast and accurate predictions for customer-focussed Intelligent Transport Systems (ITS). Traffic flow prediction has a wide range of applications, from assessing potential designs for new road layouts, reducing or removing accident hotspots, to short-term prediction of traffic congestion; see [1] [2]. Such systems also provide travellers with the potential to make informed decisions in real time, to avoid unnecessary stress, save time and, ever importantly, to reduce carbon emissions. In this research we extend historic traffic flow information to all roads in the UK with the use of spatio-temporal machine learning. This paper considers three key methods; spatio-temporal linear regression to act as a baseline model, a KNN algorithm as inspired by previous traffic flow models [3] and, a REPTree decision tree which, to the best of our knowledge, has never been utilised in traffic flow prediction.

Temporal approaches have been consistently applied in current state-of-the-art predictive models applying ARIMA, Markov Chains, Bayesian Belief Networks (BBN) and Artificial Neural Networks (ANN), primarily for short term prediction for intelligent traffic systems such as the UK's CGI-Systems [2] [4], producing a mean absolute percentage error as low as 8.6% [3]. However, these previously utilised approaches consistently lack two characteristics, (1) information regarding a road's surrounding environment and (2) information about roads that have no sensors. These additional characteristics, we believe, have potential to transform distributed traffic simulations. For example, a complete urban network can prove costly to accurately simulate at low-level [5]. However, if one were to make assumptions regarding traffic similarities based on a road's surroundings, then the number of simulations can be significantly reduced to a subset of roads unique by feature.

In the remainder of this paper we: (I) Review the most successful spatio-temporal road usage predictors published to date and assess how these predictions complement common simulation solutions; (II) Describe the process of data collection and feature selection; (III) Present a description of the data and machine learning approaches undertaken; (IV) Interrogate the results of supporting experiments before presenting concluding remarks.

II. Related Work

A. Machine Learning for Traffic Modelling

Research similar to ours was put forward by [3], who introduced an AKNN-AVL method which combines Advance K-Nearest Neighbour and a balanced binary tree data structure, resulting in a mean absolute percentage error (MAPE) of just 8.45%. This approach was put forward to overcome KNN's computational complexity and memory limitations. Our approach is not affected by these same weaknesses and in addition considers spatial variables with the use of neighbourhood features.

Research most comparable to our own involved spatio-temporal data mining (SDM) and was conducted by [2]. Here spatio-temporal features were used to predict the average daily traffic data in Hong Kong; features included shopping centre locations, home communities and car parks. Their Bayesian Belief Network (BBN) yielded an r^2 of 57.76% in some, but not all, cases of short term prediction. Our research takes the same spatial approach, but includes more variables, stricter feature selection, use of a REPTree and no forecasting.

Spatial decision trees have repeatedly performed well in traffic modelling, most notably with neighbourhood graphs based on an ID3 decision tree, which acts in a similar way to a REPTree, with the one distinct difference that the ID3

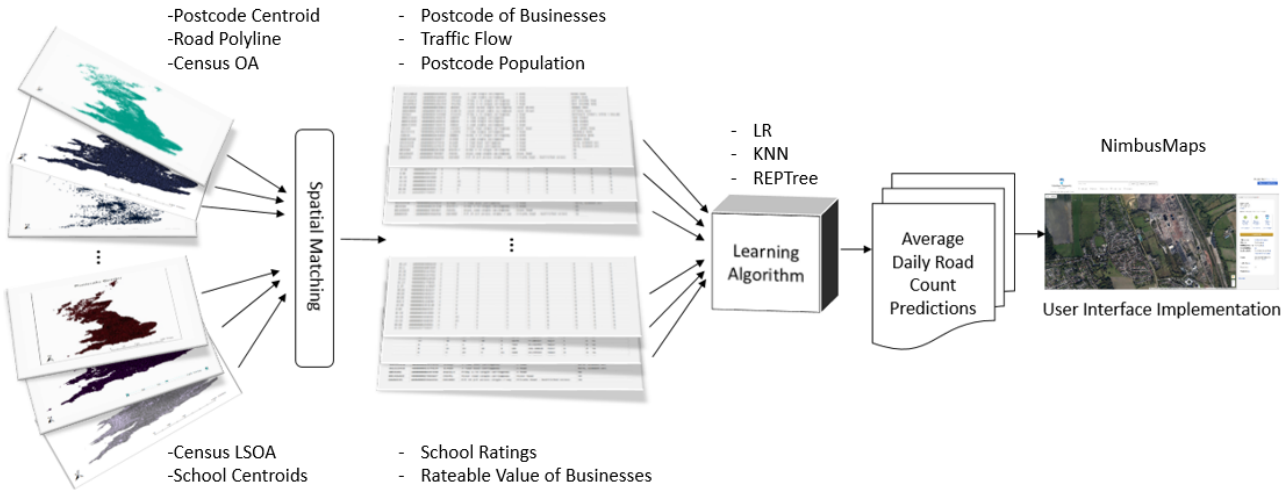

Fig. 1. Process flow for our proposed spatial analytics, which can be used as a simulation prior

algorithm does not prune [6]. Moreover, a demographic analysis undertaken by [7] included data such as income, neighbourhood population and proximity to the closest park; this work effectively considered spatial functions with non-spatial data, and generated a set of non-spatial concept hierarchies for implementation into a binary decision tree. In our research we propose that local trends (termed *features*) can be employed to determine traffic flow at a national level.

B. Techniques based on Distributed Simulation

Various techniques have been employed for the distributed simulation of real-time traffic data, including the use of cellular automata [8]. Such techniques rely on driver vehicle units in a discrete space, and aim to derive localised rules. Although fast, this approach relies on assumptions which can, in turn, produce poor accuracy. Alternatively, microscopic simulation research such as that shown by [5] and [9] utilise position, speed and acceleration of vehicles, computed periodically. A time-stepped, time-flow mechanism is used to provide highly granular simulation, with the inevitable bi-product of increased computation time. Our research aims to complement such approaches with a prior step, which may be used to increase accuracy in cellular automata-based approaches and reduce the computational complexity of microscopic simulations.

III. METHOD

A. Data Description

The open-source data discussed come in two formats; 'text' and 'shape'. Shape files are displayed in three internal file formats, .shp, .shx and .dbf, which represent the shape format, the shape index format and the attribute table respectively. Shape files can come in three different styles: points, polylines and polygons. Figure 2 gives an example in a Geographical Information System (GIS), where the 'postcode centroids' are points, the 'parcels' are polygons and the 'roads' are polylines. These shape files are input at Stage 1 of Figure 1.

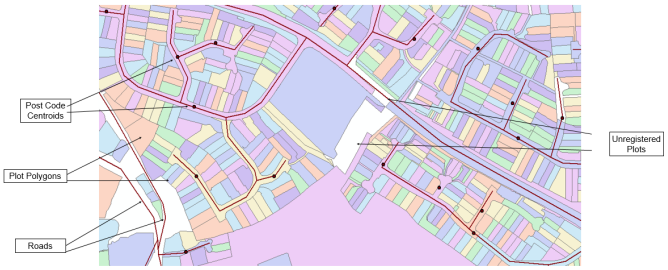

Fig. 2. An example Geographical Information System (GIS) input, showing postcode centroids, land parcels and polylines

Here the input text files (Stage 3 of Figure 1) are a set of statistical data-points related to spatial areas such as a postcode, census output area (OA) or county.

The traffic counts released by the Highways Agency England and the Department for Transport (DfT) are sourced from 8,000 road-side sensors [10]. This includes 116,615 instances of daily average traffic flow counts for a 16-year period, together with "Road Category" information ('M', 'A', 'B', 'C', 'U', where M is a motorway, A, B and C are dual or single lane roads and U is an unclassified road). Additionally, the Office of National Statistics (ONS) release a number of spatial statistics such as postcode population. Finally, the Valuation Office Agency (VOA) provided a Summary Valuation (SV) dataset of all UK commercial transactions registered. This dataset includes the land use (in 269 categories, including restaurant, shop, sports centre etc.), address and rateable value.

Feature selection was performed using best-first search, correlation and 10-fold cross validation. This ensured that, for example, a postcode with a large number of shops - and most likely a proportionally higher number of restaurants -

did not bias the final traffic flow decision tree by employing both highly correlated features. An additional benefit of this approach was that pre-selecting a subset of 21 non-redundant features improved the performance of the model calculation.

B. Training the Traffic Flow Predictor

Our approach employs spatio-temporal linear regression, a K-Nearest Neighbour algorithm (KNN) and a Reduced Error Pruning Tree (REPTree). Regressions are commonly used in industry, for example in residential valuation and stock prediction [11]. Spatio-temporal linear regression forms our baseline model; we are aware however that concerns regarding linear regressions include their uncertainty towards the cause of an ascertained relationship and their inability to capture neighbourhood interrelations and micro-variations [12] [13].

KNN forms a more robust alternative [14]. Given a set of N traffic counts $\mathbf{Y} = [y_1, ..., y_n] \; \epsilon \; R^N$, all with a set of D spatial and thematic features $\mathbf{X} = [x_1, ..., x_n] \; \epsilon \; R^{DXN}$ and an objective to predict a road's traffic count y_i' with no prior traffic flow information, but similar spatial features, the aim is to find the k-nearest neighbours in terms of common features and allocate a traffic flow accordingly. The most significant identified weakness of KNN is it having no memory.

A REPTree-based solution predicts based on a splitting criteria of information gain and prunes using reduced error [15]. This model is particularly good for reducing the error arising from variance. The root node will be the one with the highest information gain for the full training set. Then a set of child nodes will be produced with the highest information gain based on the root node's value; this continues until the minimum information gain threshold is met (0.001 in our case). We set the maximum depth to -1, the minimum number of instances to warrant a branch at 2.0, the number of folds to 10 in our iterative validation. This approach has been criticised based on its potential naivety to smaller decision trees and datasets [16]. This is true to an extent, due to the process of partitioning into three sets; training, validation and testing, which discards valuable information. However [17] shows that the accuracy of the REP method increases with the size of the tree, showing an error of $\leq$ 10% on trees of $\geq$ 100 nodes. This method has not previously been applied to spatio-temporal traffic predictions, despite the good fit to this problem domain. We also believe that this method would serve as an excellent prior for large, distributed traffic flow simulation.

C. Validation Metrics

For comparison between methods, two relative and two absolute performance metrics were applied: r^2, Root Mean Squared Error (RMSE), Mean Absolute Percentage Error and Training Time. The r^2 calculation measures the predictor's 'goodness of fit' (the model's ability to fit the test data):

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

The Relative Mean Squared Error, intuitively takes the square root of the sum of the mean squared errors:

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

MAPE is the mean absolute error, expressed as a percentage:

$$MAPE = \frac{100}{n}\left(\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{y_i}\right)$$

Finally, the training time is the time it took for 10-fold cross validation on 116,615 instances on a Windows machine with 32GB of RAM and a 3.5GHz quad-core processor.

IV. Results

On 10-fold cross validation, the linear regression produced an r^2 of 0.405, an $RMSE$ of 13279.24 and a $MAPE$ of 90.39%. 'Class legend', 'road nature', 'population' and 'households' were predicted to have the most significant dependence on traffic flow. The KNN algorithm produced an r^2 of 0.74, an $RMSE$ of 4044.92 and a $MAPE$ of 9.31%. The resulting REPtree had 115,599 nodes. The features with the highest information gain were: 'class description', 'class legend', 'primary class' and 'Rcat' (negative correlation). The REPTree produced an r^2 of 0.879, an $RMSE$ of 5385.9 and a $MAPE$ of 19.59%. It can be seen in Table 1 and Figure 3 that the REPTree produced the most effective goodness-to-fit, yielding an impressive 88.2%.

TABLE I
PERFORMANCE COMPARISON OF LR, KNN AND REPTREE-BASED SOLUTIONS

Result	Regression	KNN	REPTree
r^2	0.41	0.74	0.88
RMSE	13279.24	4044.92	5385.90
MAPE	90.39	9.13	19.59
Training	1.21	0.08	5.33

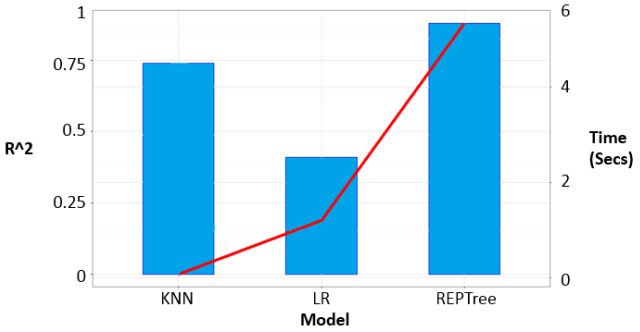

Fig. 3. Comparison of R^2 and training time between all three models

V. Implementation

The results of this research are embedded in NimbusMaps (patent pending), a property intelligence tool created by Assured Property Group. The interface, powered by GoogleMaps, provides polygons representing all available Title numbers in England and Wales. One can search by postcode or current location and then select a Title number of interest, for which ownership details, site size, flood risk, estimated residential value and traffic flow are returned. All traffic counts were split into 10 equal frequency bins, where '1' represents the bottom 10% of traffic on a single road and '10' represents the converse. This addresses the heavy skew towards low traffic flows, we also found that this was more intuitive for clients. To illustrate the tool's use, consider a fast food company with drive-through restaurants interested in developing a new outlet. The company will have specific criteria to ensure that their business is profitable, which will include traffic flow, population density etc. Using NimbusMaps the company can find all properties that sit on a road which meets these criteria.

VI. The Impact on Traffic Simulation

We believe this approach has a significant role to play in real-time distributed traffic simulation, including: (1) Its ability to decrease the size of large networks before simulations take place, to reduce sub-network load balancing and inter-process communication management; (2) Potential to extend simulations for roads with no sensors; (3) The use of large, open-source data sets and spatial machine learning to augment existing state-of-the-art techniques.

In [18], two methods for the division of road traffic networks for heterogeneous clusters are presented, which is typical amongst high-powered, distributed, traffic simulations. Additionally, [19] introduced an agent-based architecture proposed for simulating common concerns linked to urbanisation, including congestion, collisions and high levels of emissions; the research is demonstrated on a small sample of roads in Singapore. We believe that both experiments would benefit from our approach as it utilises an initial step of identifying common features on a single stretch of road in non-spatially common networks. As a result, larger networks, cities, or even countries could potentially be simulated.

VII. Conclusions

In this paper we have: (1) deployed spatial matching algorithms to combine features; (2) trained a set of traffic flow data-points on over 8,000 roads across the UK; (3) extended traditional traffic flow predictions to include spatial, temporal and neighbourhood features. Validation reports a 88.2% accuracy with the REPTree algorithm. In addition we (4) discussed the benefit of introducing spatio-temporal machine learning as a prior to real-time distributed traffic simulation and (5) implemented this research into a real estate decision engine.

Planned extensions include the use of the Department for Transport's API, whose data is updated every 10 minutes, to predict traffic counts throughout the day. This additional data will similarly be integrated into APG's NimbusMap user interface, to support consumer centralized estate management.

VIII. Acknowledgements

We thank the Engineering and Physical Sciences Research Council (EPSRC) Centre for Doctoral Training in Urban Science (EP/L016400/1) for their support.

References

[1] H. Yin, S. Wong, J. Xu, and C. Wong, "Urban traffic flow prediction using a fuzzy-neural approach," *Transportation Research Part C: Emerging Technologies*, vol. 10, no. 2, pp. 85 – 98, 2002.

[2] S. Sun, C. Zhang, and G. Yu, "A bayesian network approach to traffic flow forecasting," *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, no. 1, pp. 124–132, March 2006.

[3] J. WANG, P. SHANG, and X. ZHAO, "A new traffic speed forecasting method based on bi-pattern recognition," *Fluctuation and Noise Letters*, vol. 10, no. 01, pp. 59–75, 2011. [Online]. Available: http://www.worldscientific.com/doi/abs/10.1142/S0219477511000405

[4] 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*, pp. 865–873, April 2015.

[5] P. Gonnet, "A queue-based distributed traffic micro-simulation," Citeseer. 2001.

[6] I. S. Sitanggang, R. Yaakob, N. Mustapha, and A. A. B. Nuruddin, "An extended id3 decision tree algorithm for spatial data," in *Spatial Data Mining and Geographical Knowledge Services (ICSDM), 2011 IEEE International Conference on*, June 2011, pp. 48–53.

[7] K. Koperski, J. Han, and N. Stefanovic, "An efficient two-step method for classification of spatial data," in *proceedings of International Symposium on Spatial Data Handling (SDH'98)*, 1998, pp. 45–54.

[8] J. Vasic and H. J. Ruskin, "Cellular automata simulation of traffic including cars and bicycles," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 8, pp. 2720–2729, 2012.

[9] K. Nagel and M. Rickert, "Parallel implementation of the transims micro-simulation," *Parallel Computing*, vol. 27, no. 12, 2001.

[10] DepartmentForTransport, "Statistics at DfT. 2016," *https://www.gov.uk/government/organisations/department-for-transport/about/statistics*, 2016.

[11] P. Guhathakurta, *Long-range monsoon rainfall prediction of 2005 for the districts and sub-division Kerala with artificial neural network*. Current Science, 90,773-779, 2005.

[12] D. V. Budescu, "Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression." *Psychological Bulletin 114.3*, 1993.

[13] R. D. Yaro and R. J. Raymond, "State planning in the northeast," *Land Lines: July 2000, Volume 12, Number 4*, 2000.

[14] S. B. Imandoust and M. Bolandraftar, "Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background," *International Journal of Engineering Research and Applications*, vol. 3, no. 5, pp. 605–610, 2013.

[15] M. Zontul, F. Aydin, G. Dogan, S. Sener, and O. Kaynar, "Wind speed forecasting using reptree and bagging methods in kirklareli-turkey," Journal of Theoretical and Applied Information Technology.2013.

[16] R. A. Rinkal Patel, "A reduced error pruning technique for improving accuracy of decision tree learning," International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, 2014.

[17] T. M. Mitchell, *Machine Learning*, 1st ed. New York, NY, USA: McGraw-Hill, Inc., 1997.

[18] T. Potuzak, "Division of road traffic network for distributed simulation performed on heterogeneous clusters," in *Engineering of Computer Based Systems (ECBS), 2012 IEEE 19th International Conference and Workshops on*, April 2012, pp. 117–125.

[19] D. Zehe, A. Knoll, W. Cai, and H. Aydt, "{SEMSim} cloud service: Large-scale urban systems simulation in the cloud," *Simulation Modelling Practice and Theory*, vol. 58, Part 2, pp. 157 – 171, 2015, special issue on Cloud Simulation.