#171

THE SYNTACTIC SIGNATURE OF STARBUCKS' LOCATIONS

Towards a machine-learning approach to location decision-making

JOÃO PINELO SILVA
Department of Architecture & Interior Design, University of Bahrain, Kingdom of Bahrain
jpinelo@uob.edu.bh

ABSTRACT

The space syntax Theory of Natural Movement postulates that everything else being equal, land use selects their location based on the asymmetry of accessibility created by the configuration of the street network. In this article, I test the hypothesis whether configurational (syntactic) properties of an urban street network are relevant for the location of land use. If syntactic features are relevant, then land use types may have a 'syntactic signature'. To identify this blueprint, I apply machine-learning techniques to datasets of syntactic measures, for ten business types. The results are ten models, each with the syntactic blueprint of a type of business. The models are used to predict the existence, or not, of such businesses in segments of the map of London. The performance of the models varies, with fifty per cent reaching statistical significance, including one with 'good' prediction ability. The models for Starbucks coffee shops and solicitor's offices have the strongest prediction ability. The exploratory exercise demonstrates the potential of the machine-learning method Random Forests, when supervised and individually applied to a business activity, to identify the syntactic signature of business types. Such models can be used during planning and design and on location studies. The results strengthen the candidacy of syntactic measurements to location-decision-making. Moreover, they reinforce the theory of Natural Movement.

KEYWORDS

Space Syntax, Land Use, Machine-learning, Random Forest, Spatial Configuration, Urban Planning.

1. INTRODUCTION

One of the toughest challenges of urban development lies between the domains of urban planning and urban design. Practitioners are asked to guarantee that the design is appropriate for the planned occupation. The planner envisages land use, densities, occupancy, and often accepts a layout of the street network. The design of streets and buildings follows, to match the vision. Energy, water, communications, sewage, stormwater, among other infrastructures such as buildings and streetscape, create conditions for the intended land use, expected occupants, and expected traffic. Such arrangements set a stage, but the actors, such as businesses or families might or might not settle in. Understandably, occupancy rates are the first measure of the success of urban development, and demand for a location stimulates the rise in the value of property prices.

The success of a project depends on the suitability of all aspects. Therefore, creating conditions for the planned occupation involves designing with congruence among all infrastructure (Pinelo Silva, 2010). In this article, I discuss the matching of street configuration to land use. This relationship is significant because a mismatch is likely to impair the outcome of the project, in the extreme causing it to fail.

In this exploratory approach, I demonstrate how information extracted from the spatial configuration can be used to inform the planner of the likelihood that an intended specific type of business will locate on a site. In other words, the method can inform the planner of the probability of a match between a type of land use and a place. This knowledge would have advantages. The planners would know, from the design stage, what land use, realistically, they can aim at attracting. It would also permit the planners and designers to iterate the design to maximise the likelihood of success in attracting the desired land use which is important to provide the appropriate design and infrastructure. Furthermore, evidence-based backup of the project might make it more attractive for investors. Space syntax methodologies are already used by practitioners for iterative design with spatial analysis (Karimi, 2012). The work I report here has the potential to extend such approach in depth and facilitate reaching more tangible outputs. I follow the gist of a previous study by Pinelo Silva (Pinelo Silva, 2006), albeit with an entirely different and more robust approach.

The founders of configurational research present the natural movement theory (Hillier, B. Penn, A., Hanson, J., Grajewski, T., Xu, 1993) as the rationale behind configurational models. The theory postulates that other things being equal, the asymmetry of the spatial network creates a hierarchy of accessibility that influences businesses on the selection of location for premises. Contingent on their dependence on footfall, businesses pick locations based on network characteristics such as connectivity, integration and choice at different radii. For example, one business type might benefit from both high global and local integration, while another from great global integration, and low local integration. I use machine learning techniques to identify the pattern of location of ten business types based on the syntactic properties of their current locations in London, UK.

If the natural movement theory is correct, and land use location is dependent on syntactic properties of the configuration, then it should be possible to reverse-engineer the process and find out which levels of each syntactic features a particular land use locates on. In other words, it should be possible to find out the syntactic signature (blueprint) of a business type. To date, there is (to my knowledge) no documentation of the use of syntactic properties in the location decision-making procedures of businesses. Therefore, the process is not literally reverse-engineering, but the uncovering of a hidden order.

Statistical learning plays a key role in science, finance and the industry, often intersecting with other areas such as engineering (Hastie, Tibshirani, & Friedman, 2009b). It is used to ' learn' or identify patterns in data, leading to prediction, identification or estimation of outcomes. I, therefore, apply machine-learning on an attempt to learn the syntactic properties that are common to several instances of each type of business, for ten business types. The syntactic characteristics that are common among all instances of a business type are its syntactic signature, or blueprint.

In a typical machine learning exercise, the dataset is a compilation of the features which are thought to be most promising to inform the classification problem. In the case of business location, this would normally include population data such as density and income, for example. The challenge I take on here is different in the sense that I purposefully limit the dataset to configurational properties - space syntax measures of the segment map of the street network.

2. DATASETS AND METHODS

For this exploratory study, I selected activities that typically have small facilities and are nimble and therefore relocate relatively easily, and are dependent on foot traffic (such as cafes or hair salons) and others that sit on the other end of the spectrum (such as hospitals and primary education). Other business types may lay somewhere between the two extremes. The variety was thought to perhaps lead the way to some classification of land use types as a function of their dependence on configurational properties.

I used the UK's Companies House business database (download 28/02/2016), which contains the registered street address for all businesses listed. I selected the following business types

based on geographic location (Greater London) and on the Standard Industrial Classification of Economic Activities (SIC): Unlicensed restaurants and cafes (SIC code: 56.10/2), hairdressing and other beauty treatment (96.02), hospital activities (86.10/1), specialist medical practice activities (86.22), primary education (85.2), public houses and bars (56.30/2), solicitors (69.10/2) and travel agencies (79.11). It is common practice that some businesses do not operate at their registered address. I, therefore, added two more kinds, not of business types, but branches of brands: Starbucks (the coffee shop) and Waterstones (the bookstore). Both Starbucks and Waterstones make the addresses of their branches available on their websites, from where they were retrieved and stored in a database. The addresses were filtered for removal of branches inside shopping centres as these are not located directly on the street and therefore do not have direct syntactic properties. The addresses were geocoded, and point events created using Google Maps API in Quantum GIS (QGIS). After geolocating, businesses for which addresses were not found, were excluded from the dataset. The events were overlayed on the segment map of London (Source: Space Syntax Ltd), which had been previously analysed with UCL Depthmap X. After this operation, businesses which were located further than 50 meters from the closest segment were eliminated from the dataset. The representativeness (number of events) of each business on the dataset as analysed below are: cafes- 1,669; solicitors- 1,089; hairdressing- 2,033; hospitals-1,208; pubs-973; specialist medical pracice-953 ;primary eductaion-392; travel agencies- 1,059; Starbucks- 157; Waterstones- 26.

The segment map, where each segment was classified as containing or not an event, was imported into R (R Development Core Team, 2016) for analysis through machine-learning algorithms. Overall, the map dataset contained 455,928 segments, with nine variables. The variables, also called modelling features, were chosen based on the two foremost types of measures, often found to correlate with social and economic phenomena (Hillier, B. Penn, A., Hanson, J., Grajewski, T., Xu, 1993; Lerman, Rofè, & Omer, 2014; Penn, Hillier, Banister, & Xu, 1998). The measure Integration represents the potential for movement towards a place, while the measure Choice represents movement through a place. Each measure is represented by four radii: n (the whole system), 3200, 1600 and 800 meters (approximately 5, 10 and 15 walking minutes), plus one categorical variable that indicates if the segment has each business type or not. The existence of the latter variable can be used to develop a 'supervised' (explained below) prediction model (learner). The learner that later can be applied to predict the location of businesses based, solely, on the eight syntactic features. The data was preprocessed by normalising the four variables of the measure Choice, via logarithmic transformation.

2.1 MACHINE-LEARNING

In this article, I aim at demonstrating how statistical learning can be used to predict if a particular location, whether existing or being planned/designed, is likely to have the potential to accommodate a specific type of business. The aim is to predict the outcome of the question: Is this site syntactically similar to other Starbucks' sites? To which the answer might be either 'yes' or 'no'. The first meaning that the location has syntactic features that are similar to the current locations of Starbucks shops, and therefore has the potential to host a Starbucks shop; and the latter meaning the opposite. In this case, there are two possible outcomes (two classes): i.e. 'Starbucks' and 'no-Starbucks', therefore the problem is one of binary classification.

As in a typical machine-learning scenario, I start with an outcome variable. In this case, a categorical variable of the existence or not of one business type for each map segment. Since the method is based on the availability of known outcomes, this is a supervised learning process of classification. (In comparison, an unsupervised process would classify into classes which were not previously defined.)

A fundamental part of a model is a measure of its performance. Performance should not be measured on the same dataset that was used to make the model, as this would not reflect the ability of the model to predict on new/different data. To have a testing set to measure the performance of modelling output, I randomly divide the initial data for each business into two independent datasets. The training set, with 75% of the events; and the testing set, with 25% of events. The split is done via random subsampling, without replacement. Therefore, independent

training and testing sets are created for each business type. The training set is used to 'observe' (learn) the outcome based on features, the syntactic values. Based on this 'observation', I build the prediction model (or learner). The learner is later used to predict outcomes on the testing set. Since the real classification for the testing set is known, it is compared to the modelled outcomes. I simple terms, the delta between the two reflects the performance of the model.

On an exercise of binary classification, the outcome is two-fold: event belongs to class, or event does not belong to class, such as in 'Starbucks' or 'no-Starbucks'. When assessing the performance of the output of the binary classification, the results can be grouped into four types. True positive = correctly identified as belonging to the class; false positive = incorrectly identified as belonging to the class; true negative = correctly rejected as not belonging to the class; false negative = incorrectly rejected as not belonging to the class.

2.2 RANDOM FORESTS

I used the modelling technique Random Forests (Breiman, 2001) applied to the dataset described above. In the first step of the process, the model is created by using decision trees (Hastie, Tibshirani, & Friedman, 2009a) to 'learn' the syntactic properties of each business type, based on the current classification. On the second step, the learner (model), is applied to classification and its performance is tested.

Random Forests (RF) yield accurate models and are robust to overfitting (Hastie et al., 2009b). The RF model is based on an aggregate of decision trees. By itself, each decision tree is fast but not accurate. To improve the accuracy, RF fits several decision trees to form a model. I used the default value for the hyper-paramenter that establishes how many trees are used, which is 500. Each decision tree uses bootstrap aggregation (bagging), meaning that it is fit to a bootstrap sample of the dataset. Furthermore, the variables are re-sampled at each split (tree branch). The number of randomly selected variables used at each split is defined by the hyper-parameter 'mtry'. Since there were eight variables to train the model (ninth variable is the classification), the default was the creation of 3 'mtry's, at 2, 4 and 6 variables. To maximize the abiity of the model to learn, I induced seven 'mtry's, at 2, 3, 4, 5, 6 and 7. The software simulates learning for each of the 'mtry's and chooses the one with the best performance to create the model.

I used the R packages caret (Kuhn et al., 2012)(version 6.0-76) as well as the package ranger (Wright & Ziegler, 2017) (version 0.7.0) (results reported). I used the latter for speed and convenience since the results were similar to the ones obtained with the randomForest package (randomForest (Liaw & Wiener, 2002)(version 4.6-12)). I also used the packages: rpart (Therneau, Atkinson, Ripley, & Ripley, 2015)(version 4.1-10) and ROCR (Sing, Sander, Beerenwinkel, & Lengauer, 2005) (version 1.0-7) to calculate the probabilities and plot the ROC, with similar results to the previously used packages. I used default settings for all parameters except for 'mtry', as described above.

On a first attempt to find the syntactic signatures for all ten business types, I applied one multiclass model to classify all land use types at once. Such approach resulted in very poor prediction ability. Such performance might be due to some street segments having several different events, or because some business types do not seem to have a strong signature and predict well, as the results shown below suggest.

3. RESULTS

Below, I report the results of the models when applied to the testing datasets, which are independent of the training datasets. Therefore, the results discussed and summarised in Table 1 document the ability of the models to make predictions in different datasets, and not how well they predicted on the training data, where models typically show higher performance.

Regarding the overall performance of the model, the simplest measure is Accuracy - the overall proportion of samples correctly predicted by the model. For Starbucks, this was 75.5%, with a p-value of 0.001 at the 95% Confidence Interval (CI). However, Accuracy is a simple measure because it does not account for the possible imbalance of frequency between classes. An

alternative measure, which takes this phenomenon into account, is Kappa. For Starbucks, the Kappa is 0.51, and for solicitors is 0.63. (For reference, values above 0.3 are generally considered as acceptable.)

Other measures of performance can be used to understand specific aspects of the predictive ability of the model. Sensitivity (True Positive Rate, or Recall) is a measure of the proportion of true positives. For Starbucks, Sensitivity is 0.72. Specificity (True Negative Rate) is a measure of the percentage of true negatives. For Starbucks, Specificity is 0.79. The predictions of the existence of a Starbucks shop is correct approximately 70% of the time. The prediction of the absence of Starbucks is correct approximately 80% of the time. These and other values for all ten models are summarised in Table 1.

Business	Accuracy	Kappa	Accuracy Lower	Accuracy Upper	Accuracy PValue	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	ROC
Cafes	0.728	0.439	0.689	0.766	0.000	0.823	0.609	0.728	0.730	0.691
Hair & Beauty	0.653	0.289	0.615	0.689	0.000	0.745	0.540	0.664	0.634	0.605
Medical	0.655	0.305	0.599	0.708	0.000	0.673	0.633	0.693	0.611	0.630
Solicitor	0.824	0.633	0.781	0.862	0.000	0.916	0.703	0.802	0.865	0.734
Pr. Education	0.620	0.211	0.527	0.707	0.116	0.735	0.472	0.641	0.581	0.656
Pubs	0.772	0.522	0.722	0.818	0.000	0.852	0.659	0.780	0.759	0.775
Travel Ag.	0.733	0.447	0.683	0.779	0.000	0.808	0.633	0.748	0.710	0.746
Hospital	0.574	0.137	0.523	0.624	0.083	0.655	0.480	0.595	0.545	0.557
Starbucks	0.755	0.510	0.617	0.862	0.001	0.724	0.792	0.808	0.704	0.810
Waterstones	0.500	0.000	0.157	0.843	0.637	0.250	0.750	0.500	0.500	0.500

Table 1 - Summary of performance statistics. Accuracy Lower and Upper, as well as p-value refer to 95% Confidence Interval (CI). ROC refers to the area under the ROC curve (AUC).

Other, finer, measures of performance are: Positive predictive value - the probability that there is Starbucks when the model predicts Starbucks; or what fraction of the positive tests the model got right (what portion of predicted Starbucks actually have Starbucks). Negative predictive value - the probability that there is no Starbucks when the model predict no-Starbucks. These measures for all models are available in Table 1.

The trade-off between Sensitivity and Specificity can be visualised on a Receiving Operator Characteristic (ROC) curve (see Figure 1). The area under the curve (AUC) can be used to quantify the quality of the model, where 1 represents the perfect model (no compromise is needed in one measure to improve the performance on the other measure). Typically, values over 0.7 are considered as fair models; over 0.8 are good; over 0.9 are excellent. Starbucks' area under the ROC curve is 0.81.

Another successful model is that of solicitors, with an accuracy of 0.82 (82%) (p-value < 0.001 at 95% CI), sensitivity of 0.92 and specificity of 0.7; with an AUC of 0.73. This model consists of a robust sample with 1,089 events. All models except primary education and Waterstones reach statistical significance with p-values < 0.05 (95% CI). Waterstones has an extremely small sample and is included mostly for investigative and demonstrative purposes. Nonetheless, it represents a worthless model. Note for example it's AUC of 0.5, indicating the absence of the ability to make predictions with success beyond random (50% for a binary classification such as this one). The results are significant since the only variables used in these models are network measures of centrality, while in reality, location studies tipically include socio-demograpfic ones. Furthermore, the measure of Positive Predictive Value shouldn't be expected to perform too well, because the fact that a segment has the right chracteristics for a business does not imply that a facility already exists there.

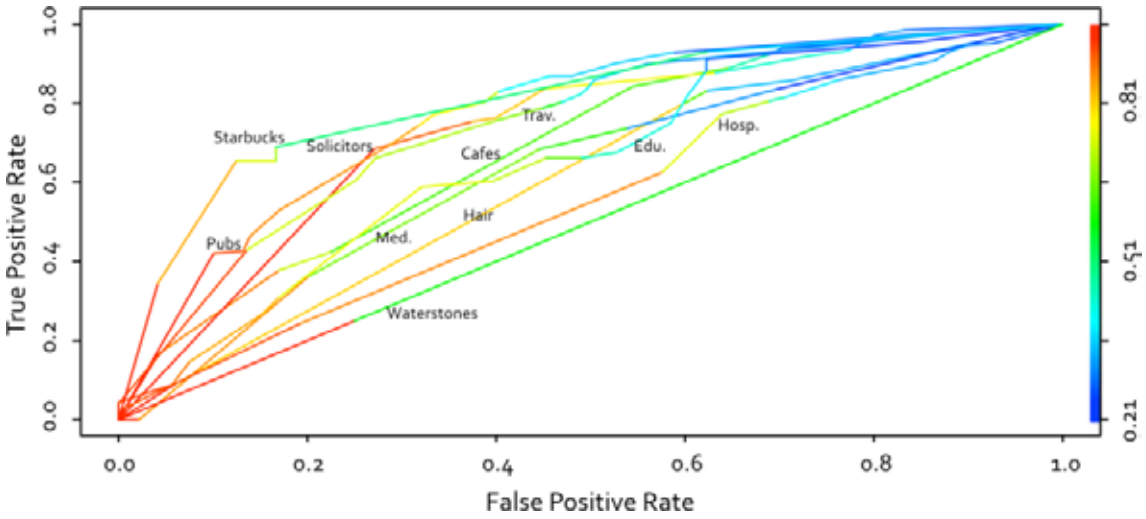

Figure 1 - Illustration of the ROC curves of the ten models. The curve of an ideal model would follow the y-axis from 0 to 1 and then the x-axis from 0 to 1. Starbucks' curve is the one that comes closer to the top let corner of the plot, and therefore have the larger area under the curve, meaning that it has the best tradeoff between sensitivity and specificity.

Further studies could perhaps investigate the differentiation between the strength of the syntactic signature of a type of business and the ability of models to predict their locations. For example, the model for the location of hospitals performs poorly, with and accuracy 0.57 (just above random). However, this should not be read as implying that, at the time of building, the hospitals did not take the configuration into consideration, since the later growth of the urban network is likely to have changed their relative accessibility. Nonetheless, the analysis seems to suggest a hierarchy of dependence on the configuration. Though it would be speculative to confirm the hierarchy itself based on this study alone, such hierarchy might be a useful way for planners to study and classify location dependency.

4. CONCLUSION

In this paper, I test the hypothesis whether configurational properties of an urban street network are relevant for the location of land use. I apply random forest modelling to a dataset of eight syntactic measures to predict the likelihood of location of a particular type of business on street segments. The supervised method was individually applied to each land use. Five of the ten models attempted have a prediction performance above 70%, statistically significant at the 95% CI. The models for the prediction of the locations of Starbucks street coffee shops and of solicitors' offices, in London, were the top performers, with an accuracy of 76% and 82%, respectively, on the testing set (simulating a real-world scenario).

Based on the results, I reject the null hypothesis - that there isn't a relationship between land use location and the syntactic properties of the configuration. Therefore, it seems possible, to some extent, to build a learner for some activities, identifying their syntactic signature. Such blueprints could be used to predict the potential of a street segment to host a business. Furthermore, this can be done since the early planning and design stages, allowing for measurable, outcome-based, design iterations.

Epistemologically, a possible interpretation of the results is that they seem to support the validity of the Natural Movement theory. A clear classification of businesses/facilities that are predictable would be too speculative at this stage, but may perhaps constitute an interesting follow up on the subject. Perhaps a classification of business types or brands based on their dependence on syntactic properties could be helpful to the field of planning, in particular for the branch of spatial decision-making.

ACKNOWLEDGEMENTS

I should thank Space Syntax Ltd. for granting access to the segment map of London for the study.

SUPPORTING MATERIALS

The machine learning process is comprehensively documented, including R computer code and both analytical and final outputs, and is available upon request. I make two documents available: a PDF file (approximately 200 pages), or an execution-able R code as Rmd file with all datasets. The documents make all the technical details available, beyond the most relevant, which are presented in the manuscript.

REFERENCES

Breiman, L. (2001). Random forest. *Machine Learning*, 45, 5–32. http://doi.org/10.1017/CBO9781107415324.004

Hastie, T., Tibshirani, R., & Friedman, J. (2009a). Random Forests. In *The Elements of Statistical Learning* (pp. 587–604). http://doi.org/10.1007/b94608

Hastie, T., Tibshirani, R., & Friedman, J. (2009b). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer. http://doi.org/10.1007/978-0-387-98135-2

Hillier, B. Penn, A., Hanson, J., Grajewski, T., Xu, J. (1993). Natural movement: or, configuration and attraction in urban pedestrian movement. *Environment and Planning B: Planning and Design*, 20, 29–66.

Karimi, K. (2012). A configurational approach to analytical urban design: "Space syntax" methodology. *URBAN DESIGN International*, 17, 297–318. http://doi.org/10.1057/udi.2012.19

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., & Engelhardt, A. (2012). Caret: Classification and Regression Training. https://Cran.R-Project.Org/Package=Caret. Retrieved from https://cran.r-project.org/web/packages/caret/caret.pdf

Lerman, Y., Rofè, Y., & Omer, I. (2014). Using space syntax to model pedestrian movement in urban transportation planning. *Geographical Analysis*, 46(4), 392–410. http://doi.org/10.1111/gean.12063

Liaw, a, & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(December), 18–22. http://doi.org/10.1177/154405910408300516

Penn, A., Hillier, B., Banister, D., & Xu, J. (1998). Configurational modelling of urban movement networks. *Environment and Planning B: Planning and Design*, 25(1), 59–84. http://doi.org/10.1068/b250059

Pinelo Silva, J. (2006). Functional Clusters on Urban Topology. In J. Van Leeuwen, Jos P ¡Tim-merman (Ed.), *8th International Conference on Design & Decision Support Systems in Architecture and Urban Planning*. Eindhoven: Eindhoven University of Technology.

Pinelo Silva, J. (2010). *Towards a Spatial Congruence Theory. How spatial cognition can inform urban planning and design*. University College London.

R Development Core Team. (2016). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing Vienna Austria*, 0, {ISBN} 3-900051-07-0. http://doi.org/10.1038/sj.hdy.6800737

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCR: Visualizing classifier performance in R. *Bioinformatics*, 21(20), 3940–3941. http://doi.org/10.1093/bioinformatics/bti623

Therneau, T., Atkinson, B., Ripley, B., & Ripley, M. B. (2015). rpart: recursive partitioning and regression trees. *R Package Version* 4.1-10, https://CR. Retrieved from http://cran.ma.ic.ac.uk/web/packages/rpart/rpart.pdf%5Cnhttps://cran.r-project.org/web/packages/rpart/rpart.pdf

Wright, M. N., & Ziegler, A. (2017). ranger : A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77(1), 1–17. http://doi.org/10.18637/jss.v077.i01