This is a peer-reviewed, accepted author manuscript of the following paper : Araldi, A., Venerandi, A., & Fusco, G. (2020). Count regression and machine learning approach for zero-inflated over-dispersed count data. Application to micro-retail distribution and urban form. In O. Gervasi, B. Murgante, S. Misra, C. Garau, I. Blecic, D. Taniar, B. O. Apduhan, A. M. A. C. Rocha, E. Tarantino, C. M. Torre, & Y. Karaca (Eds.), Computational Science and Its Applications – ICCSA 2020: 20th International Conference, Proceedings (pp. 550-565). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 12252 LNCS). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/978-3-030-58811-3_40.

Count regression and machine learning approach for zero-inflated over-dispersed count data. Application to micro-retail distribution and urban form.

Alessandro Araldi[1] [0000-0002-9732-0857], Alessandro Venerandi [1][0000-0003-4887-0120] and Giovanni Fusco[2] [0000-0002-6171-5486]

[1] Université Côte-d'Azur, ESPACE, Nice, France
[2] Université Côte-d'Azur, CNRS, ESPACE, Nice, France
{alessandro.araldi,alessandro.venerandi,giovanni.fusco}
@univ-cotedazur.fr

Abstract. This paper investigates the relationship between urban form and the spatial distribution of micro-retail activities. In the last decades, several works demonstrated how configurational properties of the street network and morphological descriptors of the urban built environment are significantly related to store distribution. However, two main challenges still need to be addressed. On the one side, the combined effect of different urban form properties should be considered providing a holistic study of the urban form and its relationship to retail patterns. On the other, analytical approaches should consider the discrete, skewed and zero-inflated nature of the micro-retail distribution. To overcome these limitations, this work compares two sophisticated modelling procedure: Penalised Count Regression and Machine Learning approaches. While the former is specifically conceived to account for retail count distribution, the latter can capture non-linear behaviours in the data. The two modelling procedures are implemented on the same large dataset of street-based measures describing the urban form of the French Riviera. The outcomes of the two modelling approaches are compared in terms of prediction performance and selection frequencies of the most recurrent variables among the implemented models.

Keywords: Retail Distribution, Urban Form, Street-network Configuration, Feature Selection, Penalised Models, Machine Learning

1 Introduction

Distribution of micro-retail is one of the most studied phenomena in urban space: the presence of stores is traditionally associated with socioeconomic dynamics and attractiveness of urban spaces [1]. Understanding the relationship between retail distribution and urban form, also named 'the morphological sense of commerce' [2], might provide academics and practitioners with evidence on how urban systems work and, ultimately, nourish the discussion on how to improve quality of life in urban areas through design and planning.

In the last two decades, a large number of empirical works investigated the association between store distribution and specific aspects of urban form [3]. Within the theoretical framework of Hillier's Movement Economy Theory (MET) [4,5], the street-network represents the most extensively explored aspect of the urban environment. MET explains how the spatial configuration of public spaces influences movement patterns and indirectly the location of stores. Inspired by this theory, several works investigated the relationship between store distribution and street-network configuration in different urban and socioeconomic contexts, for example, through the metrics of integration in Space Syntax-SSx [5] and Betweenness in Multiple Centrality Assessment (MCA) [6].

While an overall agreement on the importance of street-network properties in relation to the location of stores might be found in this specific literature, several criticisms were raised by other authors. The configurational analysis did not account, in fact, for additional aspects of urban form, such as building distribution and heights, site morphology, built-up density, which might also participate to the description of the relationship between urban form and retail distribution [3]. Together with configurational approaches, researchers have been gradually introducing additional descriptors, for example, street-based urban design qualities [7], skeletal streetscape [8], street-block typologies and built-up density [9,10], and plot system [11]. Moreover, they started to investigate how the importance of each descriptor of urban form might play different roles depending on the relative morphological context: different aspects of urban form might be associated with the presence of retail in different typo-morphological regions, such as urban fabrics [8], centre-periphery [12], the extent of the urban centre or the different morphogenetic processes (spontaneous or planned) underlying urban grids [13].

Beyond the theoretical and methodological discussion underlying the identification and conceptualisation of variables of urban form, two aspects related to the modelling procedures should be highlighted. Firstly, the aforementioned proliferation of approaches and features of urban form investigated in relation to retail distribution results in a rich yet fragmented literature. Despite evidence about the individual importance of specific aspects of urban form on store distribution, we still miss an overall picture. Assessing the combined and relative importance of a large number of urban form descriptors with innovative data analysis and feature selection procedures might provide further evidence about the relative importance of each component.

Secondly, the techniques of data analysis utilised in previous works tend to be overly simplistic [14]. Indeed, while important efforts have been devoted to the conception and implementation of sophisticated computer-aided procedures for the description of different aspects of urban form, the choice and implementation of statistical and modelling procedures have received less attention [8]. Relationships are often checked through visual inspection of maps [10], simple bivariate correlation [6,12], or Multiple Linear Regressions (MLR) [13,14]. However, the assumption of normality of residuals underlying MLR is hardly met given the usually *very skewed* and *zero-inflated* distribution of number/density of stores in the spatial unit. Only very few works considered these two specificities of the retail distribution: log-transformation of the dependent

variable [15] or count regressions approaches [9] for the former, suppression of spatial units without retail for the latter [6].

Nonetheless, the absence of retail should be considered as much informative as its presence and ignoring this specific aspect would hide important evidence on the phenomena under study. The retail distribution might be explained as the combined result of a twofold generative process defining presence/absence and number of stores, each one associated with different combinations of urban form features. Zero-Inflated Negative Binomial (ZINB) technique [16] has been shown to better perform when compared to traditional MLR and count regression [8,17], with the only downside of not being able to capture non-linear behaviours in data.

To overcome the limitations stated above and provide a more robust description of the relationship between features of urban form and retail distribution, the ZINB model developed by Araldi [17] is here compared to Gradient Boosting (GB) [18]. The former is a count regression model, part of the larger family of Generalised Linear Models (GLM) providing a built-in variable selection solution. The latter is a recently developed Machine Learning (ML) algorithm, combined with a forward feature selection, which proved to be a very versatile and robust technique of data analysis based on train and test and cross-validation [19].

These two methodologies are here compared to analyse the relationship between a large set of urban features and retail distribution in the French Riviera, a large metropolitan area located in the south of France. Outcomes, both in terms of prediction performance and selected variables, are compared and discussed. Findings show that ZINB is, in general, better than canonical GB. However, a specific nested modelling architecture combining Boolean and regressive GBs approaches to model respectively absence/presence and retail count provides higher performance levels.

The paper is structured as follows: Section 2, specify the goal of the paper. Section 3 presents the methodology under investigation: we briefly describe the study area, data sources, the set of urban form descriptors and the two modelling procedures under analysis. Outcomes of the two modelling procedures are presented and compared in Section 4, focusing both on predictive potential and the selected variables. Conclusion and perspectives of future work conclude the paper.

2 Objective

As introduced in the previous section, the major challenge in this work is how to deal with zero-inflated and highly skewed distribution characterising the retail distribution along streets. ZINB regressions were shown to handle both skewness and inflation in the data distribution in the case of retail distribution modelling, better perform among other statistical regressive approaches [8,17]. Nonetheless, an important limitation still persists: count data models might not be able to detect the presence of complex non-linear interactions between the predictors and the response variable. To overcome

this problem, GB is here implemented and compared to ZINB. GB is an improved version of Random Forests (RF), a technique of data analysis based on multiple decision trees. However, while the prediction output by RF is the average of the predictions of each of such trees, the one made by GB is obtained through an iterative process that, each time, fits new decision trees to improve predictions, and thus reduce errors [18].

However, despite a high predictive capacity, two main limitations are traditionally associated with ML (and therefore GB) approaches: firstly, the interpretation of the parameters is less straightforward. Secondly, the sample size for each class or range to be predicted should roughly be similar. When this assumption is not met, the application of a canonical ML might produce biased results focusing on the prediction of the class/range with the highest number of samples [20]. Data imbalance/skewness represents a non-trivial problem that has received growing attention in the last two decades within the ML community [20,21]. To overcome these limitations, numerous solutions have been proposed in the ML literature. Three groups might be recognised [22]: i) ad-hoc modification at the data level (i.e. over/under-sampling techniques); ii) variation at the algorithm level removing the bias towards the majority class (i.e. cost-sensitive approaches); iii) hybrid approaches combining both data and algorithm modifications.

Although many efforts targeting imbalanced distributions are regularly proposed in the community, these procedures are mainly based on heuristics aimed at the improvement of the prediction performance. Moreover, they lack relevant insights/basis on the generating process(es) underlying the phenomena to be modelled, that might guide the development of systematic imbalanced learning approaches. As recently observed by Kravzczyk [22], many shortcomings in existing methods and problems still need to be addressed appropriately. Furthermore, alterations of canonical procedures might also be associated with degradation of model performances.

For these reasons, in this work, rather than focusing on procedures that could handle imbalanced models, we propose a theory/process-based solution combining canonical well-established ML approaches. Starting from the same hypothesis underlying traditional zero-inflated regression approaches, we propose to study the retail distribution with a decomposition of the original problem into a set of two sub-problems: absence/presence and amount of stores. These two aspects are investigated with canonical classificatory (Boolean) and regressive GB, both in a disconnected and combined/nested fashion. This approach allows a straightforward comparative assessment of the two modelling approaches investigated in this paper (ZINB and RF). Moreover, feature selection procedures can be easily implemented through shrinkage (or penalised) regressions [23] and Sequential Forward Selection (SFS) [24] for ZINB and RF, respectively.

The implementation of these procedures on the same study area provides important empirical evidence about specificities of these two methodological approaches and on two major challenges: on the one side, the ability to deal with zero-inflated and highly skewed distribution characterising the retail distribution along streets, on the other, the need to identify a subset of meaningful variables from a large set, characterised by high

multicollinearity. Comparing the selected variables from the two methods allows understanding whether the underlying assumptions of traditional modelling approaches might influence the results of the analysis and if non-linearity might better explain the relationship between urban spaces and retail distribution.

Both procedures are applied to a vast metropolitan region and on nine sub-regions defined for their different morphological/contextual characteristics and degrees of skewness and zero-inflation. The spatial decomposition of the models allows assessing performances and selected variables in different contexts, providing information on the scale/contextual independence/dependency of each variable and possible non-linear behaviours between descriptors of urban form and retail.

3 Methodology

3.1 Study area and data sources

In order to assess the modelling protocols under analysis, we implement them on a real case study, that of the French Riviera, an extensive metropolitan area, including 88 coastal and inland municipalities of the department of the Alpes-Maritimes, in the southern French region Provence Alpes Côte d'Azur (PACA). Six main urban centres structure the French Riviera (Fig.1). In its western part, we find the inland town of Grasse and the two coastal cities of Cannes and Antibes, counting respectively 51, 74.2, and 73.8 thousand inhabitants. Nice, with its 343 thousand inhabitants, represents the largest municipality of the French Riviera and the administrative centre of the department. The enclave of Monaco and the border city of Menton have respectively 38 and 28 thousand inhabitants. Spread around these main centres, 295 thousand people find their home in smaller cities, villages, and hamlets. With a total of more than 1 million inhabitants, the French Riviera is the seventh most populated conurbation in France.

The combination of all these elements produces a sequence of urban centres and peripheral areas of different sizes and urban forms. Considering such a variety might help to overcome the limitation of traditional works investigating only urban cores [3]. Furthermore, the high heterogeneity of urban forms present in this study area allows a more thorough assessment of the outcomes of our two modelling approaches as different zero-inflation and overdispersion of the micro-retail distributions are observed.

Two sources of data have been used in this work. The official data on retail distribution has been provided by the local Chamber of Commerce of Nice Cote-d'Azur. Urban form descriptors are based on the geographic databases (BD TOPO®, 2017) from the French National Institute of Geographical and Forest Information. Four layers of urban morphological elements have been used: building, street-network, parcel and digital terrain model-DTM. While GIS-based protocols have been implemented for computing the different descriptors of urban forms, open-source Python and R libraries were used to implement GB (scikit-learn) and ZINB models (mpath).

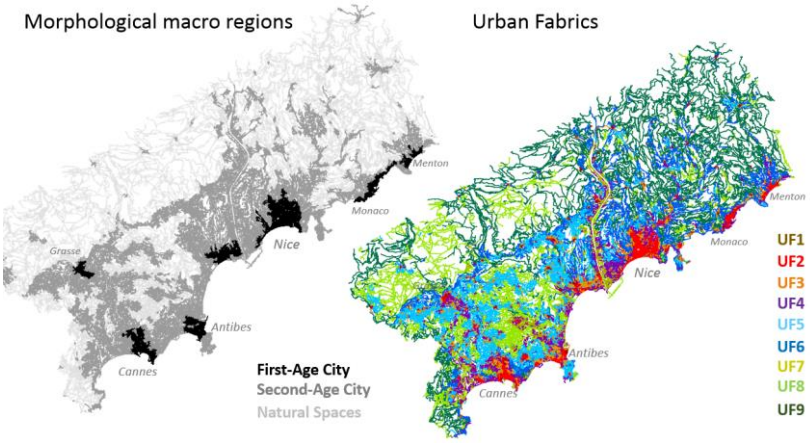

Figure 1 The study aera and the morphological regions at the macro- (First/Second-Age Cities, left) and meso-scale (the nine Urban Fabrics, right), considered in this work. Source: [8].

3.2 The variables under investigation

The spatial unit of analysis considered in this work is the street segment. It represents one of the most used spatial units and has attracted the attention of urban designers, morphologists, and geographers in the last twenty years, [25]. The street is also considered the bridging element between different theoretical backgrounds and methodological approaches [26].

The output variable was computed by assigning to each street segment, belonging to the study area under exam, the number of small stores (surface < 300 m^2) lying on it. In order to describe different aspects of urban form, several computer-aided procedures, extracted from established scientific literature, were applied to our study region. Once excluded empty streets segment,[1] our dataset had 63 thousand elements, described by more than one-hundred street-based indicators of urban form.

Such descriptors can be categorised in four main subsets. Forty indicators describe street network configurational properties and were computed through the MCA protocol [27]. Four traditional configurational indices, Reach, Straightness, Closeness and Betweenness centralities were implemented at different radii and impedances (metric and temporal, corresponding to pedestrian movement and vehicular mobility). Thirty-six indicators describe the street-network accessibility towards public squares, coastline and Anchor Stores (e.g., shopping centres, arcades etc. with an overall surface > 2000 m^2, AS), considered as influential components in the commercial fabric of cities [28]. As for the previous metrics, different radii and impedances were considered. Thirty indicators measure various aspects of urban form and were implemented through GIS procedures. These indicators describe the layout of the built form along street edges

[1] Defined as those street segment where no built-up elements are observed within a radius of 50 meters from its edges.

(also named skeletal streetscape [29]). Examples of such indicators are: façade alignment, set-back of buildings, average building height, distribution of plots, etc.

Finally, street-based contextual variables are obtained through the implementation of the Multiple Fabric Assessment procedure [30,31]. Each street segment is associated with twelve values, each describing the probability of association to nine families of urban fabrics and three morphological regions, respectively, at the neighbourhood and district level (Fig.1). These two typo-morphological partitions of the space also define the subareas where count regression and GB models were separately applied.

3.3 Modelling approaches

Based on the thematic and methodological literature previously discussed, this paper investigates the relationship of micro-retail distribution and urban form as generated by a double process (store presence and quantity), through count regression and GB. For each study (sub)region, we implement at first two couples of models exploring the aforementioned processes separately: Binomial (B) regression and GB Boolean classification for the former, Negative Binomial (NB) regression and regressive GB for the latter. The separation of these two processes requires the manipulation and separation of the original dataset that corresponds, in other terms, to an artificial introduction of expert-based knowledge in the modelling procedure. Consequently, the models of presence/absence of retail do not consider the number of stores observed along each street and their outcomes only describe streets with favourable or hostile conditions to the presence of at least one store. Contrarily, in the case of count models, outcomes describe those features of urban form which are best associated with greater/smaller numbers of stores on street segments, when/if observed. To model the combined effect of these two processes without their manual separation, specific procedures are therefore implemented.

In the GML count regression methodological framework, the Zero-Inflated Negative Binomial regression (ZINB) provides a built-in solution for the simultaneous implementation of a Binomial and a Negative Binomial regression. In such an approach, zeros are originated by two simultaneous processes: structural zeros (or true zeros) and random zeros (or false zeros). ZINB is a well-established statistical procedure already tested and implemented in several disciplines and, more recently, demonstrated to well perform when investigating retail distribution.

On the contrary, for what concerns the GB methodological framework, the absence of a specific acknowledged procedure able to consider the two processes requires us to test two different solutions. The first consists in implementing the canonical algorithm, where no difference between the two processes is made. Canonical GB is a versatile and robust technique of data analysis that combines several weak models to output a stronger overall prediction [18]. As mentioned above, the prediction made by GB is based on an iterative process that fits new decision tree models to improve predictions, and thus reduce errors, at each iteration. The minimisation of such errors is based on a loss cost function pointing in the negative gradient direction. To avoid overfitting, the

original dataset is divided in train and test subsets. First, the model is trained on a random subset of the dataset. Second, its performance is evaluated on the part of the dataset that was not used in the previous step. This procedure must be combined with the k-fold cross-validation, which consists in dividing the dataset into k folds and using each fold k−1 times as train set and once as test set to be predicted. The number of folds considered is this work is 10.

The second solution within the GB methodological framework accounts for the two processes in retail distribution (store presence and quantity) and uses several ML algorithms in a nested fashion. Firstly, to identify the best predictors of presence/absence of retail, a GB classifier is fitted in a cross-validated regime. Having obtained the model, this must be used to predict presence/absence of retail across all observations in the study area. Subsequently, a regressive GB is implemented again, in in a cross-validated regime, to predict retail count where the previous model predicted presence. The presented nested solution allows to specifically consider the two process of retail presence/absence and count without manual separation of the two subsets.

In order to deal with the redundant and highly correlated information originated by the large number of variables,[2] both GLM and GB procedures are combined with specific variable selection procedures to allow the identification of the most significant variables of urban form related to retail distribution. In the case of count regressions (B, NB and ZINB), penalised regression approaches are applied: the notion underlying this procedure is to shrink the regressive coefficients toward zero. The coefficients associated with the variables with a minor contribution to the outcomes of the model are close or equal to zero. In this way, the complexity of the model is reduced. The specific ElasticNet (Enet) procedure [32] is implemented in this work.For what concerns the ML approach, GB models (both classificatory and regressive) are all preceded by a Sequential Forward Selection (SFS). SFS identifies the best predictors of the output variable through an iterative process based on a regressor performance that adds one variable at the time until an optimal subset of features is reached [19].

The implementation of Enet and SFS procedures, allow us to identify those subsets of variables associated with retail presence/absence and count separately, within each sub-region. To compare the outcomes of all modelling solutions (ZINB, canonical GB, and nested GB), F1 scores must be computed to evaluate their performances. The F1 score measures a model's accuracy as the harmonic mean of precision and recall [33]. The former represents the number of correct positive predictions divided by the total number of all positive predictions output by the model. The latter is the number of correct positive predictions divided by the number of all samples that should have been identified as positive. F1 scores range between 1 (perfect accuracy) and 0 (worst accuracy). These performance measures describe the combined effect of the variable selection and modelling procedures within the GLM and ML approaches where specific subset of selected variables underlies each model. We will, however, remark that F1 scores are better suited to compare the zero-part of the models (presence/absence of stores) than the count part. For the latter, any predicted value differing from the observed one

[2] Which might especially affect the GLM based on the assumption of independent regressors.

will contribute equally to lowering the F1 score, regardless of the magnitude of the difference. More specific measures, like the area under the Count-REC (Regression Error Characteristic) curve [8] could be used to compare the quality of the count models. As for the GM models, F1 scores are always calculated as an average of the 10 test subsets within the k-fold procedure.

The outcomes of the variable selection procedures are finally presented in terms of selection frequency. Comparing the selected variables from each couple of models (B vs Boolean GB, zero-truncated NB vs regressive GB, ZINB vs canonical/nested GB) allows identifying the degree of similarity of the GLM and ML modelling approaches.

4 Results

3.4 Comparing model goodness of fit values

In Table 1, the outcomes of all 81 models are collected and compared. For each study (sub)region, zero rates and variances of the count part are provided. By visually inspecting Table 1, we note that higher zero-inflation is associated with lower overdispersion in non-dense sub-regions while dense urban fabrics are associated with lower zero-inflation and higher variability. F1 scores are separately evaluated for the zero and count parts allowing a direct comparison and evaluation of the different modelling procedures in relation to conditions of zero-inflation and overdispersion. Before discussing performances, we remind the reader that each model is based on different subsets of variables, since different feature selection procedures had to be used in GLM and ML. The performances showed in Table 1 are to be considered the combined result of such feature selections and modelling, implemented on the same original dataset of 105 descriptors of urban form and optimised for each sub-region under analysis.

Depending on the process(es) under analysis and judging solely from F1 scores, different outcomes can be outlined: when focusing on the presence/absence prediction of stores (zero part), both Binomial and GB Boolean models show better performances for sub-regions with more zero-inflation. F1 scores of the two modelling approaches converge for sub-regions with higher zero-inflation (less than 1,5% of variation), although traditional Binomial regressions tend to perform slightly better than GB Boolean classifiers. On the other hand, the latter tends to perform better than Binomial regressions in cases with less zero-inflation.

When modelling the number of stores (count part) with zero-truncated Negative Binomial regressions and regressive GB, the latter approach is consistently associated with higher predictive capacity, with an improvement of the F1 scores between 25% and 125%. Furthermore, stronger improvements seem to be associated with larger subsets rather than with smaller ones, which also tend to be more zero-inflated and/or overdispersed.

When the two processes of presence/absence and quantity of retail are modelled with ZINB and canonical GB, we observe that: i) F1 scores of both zero and count parts are

lower than the ones obtained through models that considered the two processes separated; ii) canonical GB performs slightly better for both zero and count parts, when considering the entire dataset; iii) in compact/planned urban fabrics, where greater numbers of stores can be found, the ZINB outperforms the canonical GB for both zero and count parts; vi) in the remaining subspaces, GB performs better in the zero part, while ZINB provides better predictions for the count part.

The proposed nested GB protocol, inherit the higher predictive performance from the separate implementation of Boolean and regressive GB. We observe that the nested GB always performs better than the canonical GB, in both zero and count parts. When comparing nested GB to ZINB, we observe that: i) F1 scores of the zero parts for the former are always greater than the ones for the latter, although the improvement decreases in those cases with stronger zero-inflation; ii) F1 scores relative to the count part of the nested GB are greater than the ones for ZINB, with the exception of those urban fabrics with grater variability of values. Finally, in UF2 (traditional planned urban fabrics with adjoining buildings), nested GB and ZINB show similar predictive performances, while, in UF4 (modern discontinuous urban fabrics with big and medium-sized buildings), ZINB outperforms the nested GB approach.

Table 1 Performance of models subdivided for morphological contexts/regions. Count variance and zero-inflation are reported in the first row. F1 scores are provided separately for the zero and count parts. The darker the red, the stronger the improvement in performance of GB approaches over GLM ones. The darker the green, the stronger the improvement in performance of regressive GLM over GB ones.

Modelled process		Dependent Variable (stores per street)	Study Region	Global	First	Second	UF1	UF2	UF3	UF4	UF5	UF6
			Variance (count part)	28,40	40,62	16,55	11,86	50,62	5,00	27,66	2,38	1,30
			% zeros	77,4	63,1	85,7	74,9	52	81,8	80,5	88,7	89,6
Separate models for micro-retail Presence/ Absence and Count	Zero Model	Binomial		0,911	0,830	0,924	0,872	0,763	0,902	0,893	0,943	0,948
		GB Boolean		0,913	0,843	0,923	0,874	0,758	0,896	0,896	0,942	0,947
		±		0,26%	1,54%	-0,07%	0,21%	-0,62%	-0,63%	0,37%	-0,12%	-0,16%
	Count Model	NB		0,170	0,123	0,239	0,195	0,110	0,285	0,155	0,399	0,484
		GB Regressive		0,366	0,224	0,408	0,312	0,148	0,452	0,218	0,585	0,618
		±		115,34%	81,64%	70,50%	59,63%	34,43%	58,77%	40,55%	46,57%	27,68%
Combined model for micro-retail Presence/ Absence and Count	Zero Part	ZINB		0,883	0,764	0,909	0,843	0,676	0,885	0,863	0,934	0,948
		GB Regressive		0,903	0,812	0,906	0,853	0,696	0,892	0,875	0,944	0,949
		±		2,25%	6,36%	-0,25%	1,16%	2,98%	0,82%	1,40%	1,11%	0,14%
		GB nested		0,913	0,843	0,923	0,874	0,758	0,896	0,896	0,942	0,947
		±		3,4%	10,4%	1,6%	3,6%	12,1%	1,2%	3,9%	0,9%	-0,1%
	Count part	ZINB		0,102	0,157	0,154	0,214	0,133	0,220	0,169	0,145	0,180
		GB Regressive		0,106	0,115	0,084	0,174	0,108	0,117	0,028	0,000	0,000
		±		4,55%	-27,04%	-45,47%	-18,81%	-18,94%	-46,85%	-83,65%	-100,00%	-100,00%
		GB nested		0,209	0,216	0,337	0,290	0,134	0,412	0,126	0,367	0,705
		±		105,8%	37,5%	119,2%	35,3%	0,7%	86,9%	-25,2%	153,8%	291,1%

To conclude, when retail absence/presence and count are modelled separately, Boolean GB and the traditional Binomial model show similar results for absence/presence, while GB outperforms zero-truncated NB in predicting the number of stores. Nested GB outperforms canonical regressive GB when the output variable shows skewed and

zero-inflated distributions, both in zero and count processes. When comparing nested GB and ZINB regressive approaches, the former outperforms the latter, especially in those cases with stronger zero-inflation. Conversely, in cases with less 0s and stronger dispersion, the two modelling approaches output similar results.

4.2 Comparing outcomes of feature selections

The attentive interpretation of regression coefficients and feature importance output by each model goes beyond the goals (and limits) of this work. In this section, we will focus instead only on checking similarities and differences between the indicators of urban form selected by the two feature selection procedures (Enet and SFS), in each model.

The number of selected indicators varies between few, in the case of non-compact peripheral regions, to many in compact urban areas (up to 30 variables, a third of the input variables). The nested GB approach displays the greatest number of selected indicators in the overall study area and in compact regions (i.e. between 18 and 30), due to the double selection procedure for each retail distribution process under analysis (presence/absence and quantity). To summarise and assess the importance of each of the chosen indicators, the frequency of their presence in the models is evaluated. Tables 2 and 3 show such frequencies for the different modelling approaches, retail distribution processes, and spatial subsets.

The first column of Table 2 reports those indicators with the highest frequencies (> 30%), across all 63 models. Among the ten most selected indicators, seven describe morphometric properties of the skeletal streetscape: Street Length, Building Coverage Ratio, Street Acclivity, Street Corridor Effect, Built-up Fragmentation, Average Building Height and Street Open Space. Local Betweenness at different scales (1,200 meters, 5 and 20 minutes) are the most selected variables among the configurational descriptors. In the next two columns of Table 2, frequencies are separately reported for the two modelling approaches, GLM and GB techniques. Five of the aforementioned indicators are found relevant for retail distribution independently by the modelling procedure (Street Length, Corridor Effect, Coverage Ratio, Built-up Fragmentation, Betweenness at 1,200m). However, we observe higher selection frequencies in the GLM approaches compared to GB ones. The former identifies a similar subset of variables for the different spatial partitions. The latter tends instead to select more diverse sets of variables for each model, seizing distinctive characteristics in each morphological subregion. Moreover, GB approaches tend to select variables with a regionalised distribution such as Slope, contextual morphometric partitions (UFs and morphological regions) and proximity to specific features (coastline and AS). They also tend to select variables describing punctual/discrete occurrences (i.e. cul-de-sac). Here again, the reason underlying these outcomes might be related to the ability of GB approaches to model non-linear relationships. In the last two columns of Table 2, frequencies are reported considering zero/count models for compact (First Age City, UF1-3) and sprawled/modernist (sub)regions (Second Age City, UF4-6), separately. We observe

how the same indicators of urban form might play different roles in the two retail distribution processes. For instance, Street Corridor Effect appears to have a relatively higher importance in defining the number of stores rather than their presence/absence, in compact contexts; however, the opposite behaviour is observed in non-compact regions. Similar presence of configurational and morphometric indicators is as influencing the retail presence/absence independently by the urban context; nonetheless, count process in compact areas seems to be more importantly influenced by morphometric streetscape descriptors than configurational ones.

Table 2 Outcomes of feature selection procedures. Selection frequencies of the most recurrent descriptors of urban form, in relation to micro-retail spatial distribution. Frequencies are here reported considering all 63 models under analysis, grouped by modeling approach (GLM/ML), Zero/ Count Parts, in Compact and Sprawled/Modernist morphological regions. Background colors identify groups of urban form descriptors: yellow- street-network configuration, light-green - skeletal streetscape, green - urban fabrics, blue - directional descriptors.

ALL (63)		Modeling approach				Compact		Sprawled/Modernist	
		Regression (27)		Machine Learning (36)		Zero Parts (18)		Zero Parts (18)	
Street Length	75%	Street Acclivity	85%	Street Length	72%	Build. Coverage Ratio	100%	Street Corridor Effect	88%
Street Corridor Effect	70%	Street Corridor Effect	81%	Street Corridor Effect	61%	Street Length	100%	Street Length	88%
Build. Coverage Ratio	67%	Build. Coverage Ratio	78%	Build. Coverage Ratio	58%	Betweenness 1200m	75%	Betweenness 1200m	75%
Street Acclivity	63%	Street Length	78%	Slope	56%	Betweenness 5min	75%	Betweenness	63%
Betweenness 1200m	60%	Betweenness 1200m	74%	Cul de sac	53%	Street Corridor Effect	75%	Street Acclivity	38%
Betweenness 5min	54%	Avg Build. Height	67%	Reach AS 300m	53%	Straightness 5min	75%	Avg Open Space	38%
Built-up Fragmentation	54%	Betweenness 5min	67%	UF9	53%	Straightness 1200m	75%	Betweenn GS	38%
Avg Open Space	48%	Built-up Fragmentation	59%	Betweenness 1200m	50%	Street Acclivity	63%	Betweenness 20min	38%
Avg Build. Height	41%	Avg Open Space	52%	Built-up Fragmentation	50%	Betweenness 20min	63%	First Age City	38%
Betweenness 20min	41%	Small Houses (<125m²)	52%	Build.Specialisation	50%	Built-up Fragmentation	63%	Built-up Fragmentation	38%
Cul de sac	40%	Straightness 5min	52%	Street Acclivity	47%	Avg. HW ratio	50%	Straightness 5min	38%
Straightness 5min	40%	Betw.AS 1200m	48%	Betw. AS 300m	47%	Avg Open Space	50%	Avg Build. Height	25%
Betw. Coast 2400m	37%	Straightness 20 min	48%	Avg Open Space	44%	Reach1200m	50%	Small Build. (250-1000m²)	25%
Build.Specialisation	35%	Betweenness 20 min	44%	Betweenness 5min	44%	Straightness 300m	50%	Build. Coverage Ratio	25%
UF9	34%	Freq Parc	44%	Clos. Coast 1200m	44%	Avg Build. Height	38%	Cul de sac	25%
Reach AS 600m	32%	Sd. Building Set-back	44%	UF1	44%	Count Parts (18)		Count Parts (18)	
Slope	32%	Straightness 600m	41%	Closeness 1200	42%	Street Corridor Effect	88%	Street Acclivity	50%
UF2	30%	Straigh. AS 1200m	41%	Reach AS 600m	42%	Street Length	88%	Betweeness 300m	50%
Betw.AS 1200m	30%	Betw. Coast 2400m	37%	Betw. Squares 300m	39%	Street Acclivity	75%	Betweenness 5min	50%
Reach AS 300m	30%	Straigh. Coast 600m	37%	Betweenness 20min	39%	Avg Build. Height	63%	Small Build. (250-1000m²)	50%
Straightness 300m	30%	Sd. Build. Height	33%	Betw. Coast 2400m	36%	Betw. Coast 2400m	63%	Street Corridor Effect	50%
Straightness 600m	30%	Straightness 1200m	33%	Closeness AS 300m	36%	Betweenness 20min	63%	Build.Specialisation	50%
Straigh. AS 1200m	30%	Reach 20 min	30%	First Age City	36%	Build. Coverage Ratio	63%	Straightness 20min	50%
		Straightness 300m	30%	Natural Spaces	36%	Cul de sac	63%	Betw. Coast 2400m	38%
		Straigh. AS 600m	30%	Second Age City	36%	Built-up Fragmentation	63%	Betw. Squares 300m	38%
				Straight. AS 300m	36%	Sd. Building Set-back	63%	Ave. Build. (1000-4000m²)	38%
				Straight.Squares 300m	36%	Straightness 600m	63%	Large Build. (>4000m²)	38%
				Clos.Squares 300m	33%	Avg. Open Space	50%	Build. Coverage Ratio	38%
				Reach Squares 600m	33%	Betw. AS 1200m	50%	First Age City	38%
				Closeness 20min	31%	Betweenness 1200m	50%	Reach AS 600m	38%
				Reach AS 1200m	31%	Betweenness 5min	50%	Second Age City	38%
				Reach Squares 300m	31%	Small Houses (<125m²)	50%	Straight. Coast 600m	38%
				Straightness 5min	31%	Closeness Coast1200m	50%	Straight. Coast 2400m	38%
				Straightness 300m	31%	Reach 20 min	50%	Street Length	38%
				UF2	31%	Reach Squares 1200m	50%	UF1	38%
				UF7	31%	Sd. Open Space	50%	UF6	38%

Street-based indicator group

Skeletal streetscape
Typo-morphologicalcontext
Street-Network Configuration
Proximity towards Coastline, Anchor Stores (AS) and Squares

In Table 3, we report the top 30 most selected features of urban form (the stronger the red, the greater the frequency of a feature). Models are divided only by considering the study (sub)regions. In this case, higher values describe greater importance of the

variable in relation to retail, independently by the modelling procedure and by the separation of the zero and count processes. Among the top 10 most selected indicators, we still find the indicators mentioned in the previous paragraph (i.e. Street Length, Building Coverage Ratio, Street Acclivity, Street Corridor Effect, Built-up Fragmentation, Average Building Height and Street Open Space, Local Betweenness at 1,200 meters, 5 and 20 minutes). Nonetheless, when considering each sub(region), different frequencies can be observed: for instance, in compact city centres (First-Age city), together with the ten aforementioned indicators, Standard Deviation of the Building Set-back and Straightness centrality at 1,200m also play important roles in the definition of retail distribution.

Table 3 Outcomes of feature selection procedures. Selection frequencies (highlighted in red) of the most recurrent descriptors of urban form, in relation to micro-retail spatial distribution, in the French Riviera (Global), in the context-based partitions at the district scale (Fist/Second Age City), and at the neighbourhood scale (UF1-6). Background colors identify groups of urban form descriptors: yellow- street-network configuration, light-green - skeletal streetscape, green - urban fabrics, blue - directional descriptors.

	Global	First Age City	Second Age City	UF1	UF2	UF3	UF4	UF5	UF6
Street Length	100%	100%	86%	71%	86%	86%	71%	57%	14%
Street Acclivity	100%	100%	71%	71%	86%	29%	57%	29%	29%
Build. Coverage Ratio	86%	100%	86%	57%	100%	86%	29%	29%	29%
Betweenness 1200m	100%	100%	57%	57%	57%	43%	57%	57%	14%
Street Corridor Effect	86%	71%	86%	57%	86%	71%	71%	29%	71%
Built-up Fragmentation	86%	100%	57%	43%	86%	29%	71%	14%	
Betweenness 5min	71%	71%	43%	29%	71%	71%	71%	57%	
Avg Building Height	86%	71%	57%	43%	43%	43%		14%	14%
Betweenness 20min	71%	57%	43%	14%	86%	71%	29%		
Avg. Open Space	57%	100%	57%	29%	29%	71%	43%	29%	14%
Straight. AS 1200m	86%	29%	43%		57%	29%	14%	14%	
Betw. Coast 2400m	43%	57%	57%	57%	43%	43%	29%		
Straightness 5min	43%	57%	43%	57%	43%	71%	29%	14%	
Straightness 600m	57%	57%	29%	29%	43%	43%		14%	
Straightness 20min	57%	29%	57%		57%	14%	14%	14%	14%
UF2	86%	43%	29%	14%				29%	43%
Cul de sac	29%	43%	57%	29%	57%	43%	43%	43%	14%
Straightness AS 600m	43%	29%	43%	29%	57%		29%	14%	14%
Small Houses (<125m²)	29%	43%	57%	29%	57%	29%		14%	
Reach20min	43%	29%	29%	43%	57%	14%	14%	29%	
Straightness Coast 600m	43%	29%	57%	29%	43%		14%		
Large Houses (125-250m²)	57%	29%	57%		14%				
Build. Specialisation	29%	29%	71%	29%	29%	57%	43%	14%	14%
Sd. Building Set-back	14%	71%	14%	29%	57%	14%	29%		
Closeness 1200m	43%	43%		14%	43%	29%	43%		
Straight. Coast 1200m	57%	43%	29%	29%					
Betweenn AS 1200m	29%	43%	57%	29%	14%	29%	29%		43%
Straightness 1200m	14%	71%	71%	43%		43%	14%		
Straightness 300m	14%	29%	43%	100%	57%		29%		
Reach 600m	43%		29%	43%	57%	14%			14%

We finally compared the sets of variables selected by the two modelling approaches implemented in each region. For each couple of models (B vs Boolean GB, NB vs regressive GB, ZINB vs canonical/nested GB), we considered the number of variables found in common and computed a Similarity Index as the harmonic mean of the rate of the shared variables. This indicator describes the degree of resemblance of the selected variables between GLM and ML approaches (thorough Enet and SFS, respectively).

The sets of selected variables tend to show greater Similarity Indexes in larger regions (Global, First/Second Age city with values between 0.31 and 0.49). Conversely, values are lower in smaller partitions (each single UF, with the exception of UF2, for

which values are still between 0.30 and 0.36). Higher similarity can also be observed for the absence/presence models between Binomial regression and Boolean GB approaches. In general, the harder the task (count vs absence/presence) and the smaller the spatial domain, the more specific the models produced by the different approaches.

5 Conclusion and discussion

In this work, we proposed a comparative study between GLM and ML approaches to explain the relationship between descriptors of urban form and number of stores, along street segments. Following previous works by Araldi [8,17], retail distribution was modelled in two ways: firstly, by applying separate models on the presence/absence and quantity of retail and, secondly, by using specific solutions able to model the two processes conjointly. To assess the two modelling procedures, implemented in these two different manners, we applied them on the same dataset, describing 105 different street-based aspects of the urban form of the French Riviera, a large coastal conurbation located in the south of France. The two modelling approaches were tested on the whole study area, but also on smaller morphological sub-regions, with different conditions of zero-inflation and overdispersion.

For what concerns model performances, similar outcomes between GLM and GB approaches were observed when modelling presence/absence; the latter proved more successful when describing the number of stores. When modelling the combined effect of the presence/absence and count processes, the canonical GB model showed lower performance compared to the ZINB model. On the contrary, the nested GB proposed in this paper proved to be a better modelling solution for dealing with the zero-inflation of retail distribution. Just like ZINB, the nested GB approach does not inject any expert knowledge in data partitioning and is not prone to survivorship biases. Nonetheless, when modelling highly skewed distribution in specific urban fabrics (UF2-4), the nested GB did not outperform ZINB.

For what concerns the outcomes of the feature selection, similarities among sets of variables were stronger in larger and central (sub)regions, while they were weaker in smaller and peripheral urban fabrics. The most recurrent variables tended to be street properties (Length and Acclivity), streetscape descriptors (Building Coverage Ratio), and aspects related to the layout of buildings (i.e. Corridor Effect, Built-up Fragmentation, Building Height and Open Space). A key role was also found to be played by local Betweenness centrality, while other configurational indices were found to be important only in specific urban contexts.

The work proposed in this paper lies the basis for more advanced comparative studies that would provide better descriptions – linear and non-linear – of the relationship between features of urban form and retail distribution. We argue that this could be very helpful to confirm or reformulate previous theories, but also to propose new ones.

In this work, we only evaluated the frequencies of the variables selected in the different models. Future work will focus on interpreting the behaviours and relative magnitudes of such variables in light of further aspects of urban morphology and concurrent urban phenomena acting on the study area.

For what concerns the approaches presented in this paper, future work might develop improvements to classic GB algorithms that would allow a better modelling of zero-inflated/skewed distributions, for example through the combination of weak (i.e. decision trees) and strong estimators (i.e. models for count data) as proposed by [34] or with multi-output modelling approaches [35]. Advanced cross-validation techniques specifically conceived for highly spatially correlated data [36] might be also considered.

Intelligibility of model results is also an issue [37]. Statistical models are easier to interpret: the signs of coefficients indicate whether regressors contribute positively or negatively to the target variable. The same cannot be said for ML approaches, and more sophisticated techniques are needed to help the understanding of model results [19]. Finally, future research might focus on the implementation of the same procedures to analyse the relationship between urban form and other urban phenomena, such as the number of traffic accidents, tweets, etc.

References

1. Chiaradia, A., Hillier, B., Schwander, C., Wedderburn, M.: Spatial Centrality, Economic Vitality/Viability. In Proc. 7th International SSx, KTH Royal Institute of Technology, pp. 16.1-16.19, Stockholm, Sweden (2009).
2. Saraiva, M.: The morphological sense of commerce: Symbioses between commercial activity and the form and structure of Portuguese medium-sized cities, PhD, Univ. of Porto, Porto (2013).
3. Saraiva, M., Marques, S.T., Pinho, P.: Vacant Shops in a Crisis Period – A Morphological Analysis in Portuguese Medium-Sized Cities, Plan. Practice & Res., 34(3), 255-287 (2019).
4. Hillier, B.: Space is the machine. Cambridge University Press, Cambridge (1996).
5. Hillier, B., Iida, S.: Network and Psychological Effects in Urban Movement, in: A. Cohn & D. Mark (Eds), Spatial Information Theory, pp. 475–490, Springer, Berlin (2005).
6. Porta, S., Strano, E., Iacoviello, V., Messora, R., Latora, V., Cardillo, A., Scellato, S.: Street centrality and densities of retail and services in Bologna, Italy. Environment and Planning B: Planning and design, 36(3), 450-465 (2009).
7. Remali, A. M., Porta, S., Romice, O.: Correlating street quality, street life and street centrality in Tripoli, Libya. The past, present and future of high streets, 104-129 (2014).
8. Araldi, A.: Retail distribution and urban form: Street-based models for the French Riviera, Doctoral dissertation, Université Côte d'Azur, Nice (2019).
9. Ye, Y., Li, D., Liu, X.: How block density and typology affect urban vitality: An exploratory analysis in Shenzhen, China. Urban Geography, 39(4), 631-652 (2018).
10. Joosten, V., Van Nes, A.: How block types influence the natural movement economic process: Micro-spatial conditions on the dispersal of shops and Café in Berlin. In 5th International SSx, Delft, The Netherlands (13), (2005).
11. Bobkova, E., Marcus, L., Berghauser Pont, M., Stavroulaki, I., Bolin, D.: Structure of plot systems and economic activity in cities: Linking plot types to retail and food services in London, Amsterdam and Stockholm. Urban Science, 3(3), 66 (2019).
12. Saraiva, M., Pinho, P.: Spatial modelling of commercial spaces in medium-sized cities. GeoJournal, 82(3), 433-454, (2017).
13. Omer, I., Goldblatt, R.: Spatial patterns of retail activity and street network structure in new and traditional Israeli cities. Urban Geography, 37(4), 629-649, (2016).

14. Sevtsuk, A.: Path and place: a study of urban geometry and retail activity in Cambridge and Somerville, Doctoral dissertation, Massachusetts Institute of Technology (2010).
15. Cutini, V. Centrality and land use: three case studies on the configurational hypothesis. Cybergeo: European Journal of Geography (2001).
16. Greene, W.H.: Accounting for excess zeros and sample selection in Poisson and negative binomial regression models (1994). https://archive.nyu.edu/bitstream/2451/26263/2/94-10.pdf
17. Araldi, A.: Towards an integrated methodology for model and variable selection using count data. Application to micro-retail distribution. Urban Science, 4(2), 21 (2020).
18. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232 (2001).
19. Venerandi, A., Fusco, G., Tettamanzi, A., Emsellem, D.: A machine learning approach to study the relationship between features of the urban environment and street value. Urban Science, 3(3), 100 (2019).
20. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on knowledge and data engineering, 21(9), 1263-1284 (2009).
21. Japkowicz, N., Stephen, S.: The class imbalance problem: A systematic study. Intelligent data analysis, 6(5), 429-449 (2002).
22. Krawczyk, B.: Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence, 5(4), 221–232 (2016).
23. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288 (1996).
24. Raschka, S.: MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. Journal of open source software, 3(24), 638 (2018).
25. Fleury, A.: La rue: un objet géographique? Tracés. Revue de Sci. Hum., (5), 33-44 (2004).
26. Kropf, K.: Bridging configurational and urban tissue analysis. Proc.11th Space Syntax Symposium, Lisbon 165.1-13, (2017).
27. Sevtsuk, A., Mekonnen, M. Urban network analysis. Revue Int. de Géom., 287, 305 (2012).
28. Brown, S. Retail location at the micro-scale. Service Ind. Journal, 14(4), 542-576 (1994).
29. Harvey, C., Aultman-Hall, L., Troy, A., Hurley, S.E.: Streetscape skeleton measurement and classification. Env. and Plan. B: Urban Analytics and City Science, 44(4), 668-692 (2017).
30. Araldi, A., Fusco G.: From the built environment along the street to the metropolitan region. Human scale approach in urban fabric analysis, Environment and Planning B: Urban Analytics and City Science, 46(7), 1243–1263 (2019).
31. Fusco, G., Araldi, A.: The Nine Forms of the French Riviera: Classifying Urban Fabrics from the Pedestrian Perspective. In 24th ISUF International Conference. Book of Papers (1313- 1325). Editorial Universitat Politècnica de València (2017).
32. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology), 67(2), 301-320 (2005).
33. Van Rijsbergen, C.J.: Information retrieval. (1979).
34. Garcia-Marti, I., Zurita-Milla, R., Swart, A. Modelling tick bite risk by combining random forests and count data regression models. PloS one, 14(12), (2019).
35. Roberts, D. R., Bahn, V., Ciuti, S., et al.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8), 913-929 (2017).
36. Borchani, H., Varando, G., Bielza, C., Larrañaga, P.: A survey on multi-output regression. Wiley Interdisc. Reviews: Data Mining and Knowledge Discovery, 5(5), 216-233 (2015).
37. Hofman, J.M., Sharma, A., Watts, D.J.: Prediction and explanation in social systems. Science, 355, 486–488 (2017).