Application of Statistical Inferencing Techniques in Building Inventory Compilation

Pooya Sarabandi[1], **Anne S. Kiremidjian**[1], **Ronald T. Eguchi**[2]
[1]Department of Civil and Environmental Engineering, Stanford University, USA.
[2]ImageCat. Inc., Long Beach, CA, USA.

ABSTRACT

In this paper a methodology for inferring engineering attributes of the built-environment, i.e. the structural type and occupancy type of buildings, from 3-D building models is formulated. The multinomial logistic model is utilized in this paper in order to compute the marginal probability distribution of class-memberships. This model incorporates and supports categorical as well as quantitative data. The explanatory variables used as the input parameters to the statistical model, are selected such that they can directly be derived from the 3-D models reconstructed from the satellite imagery. Two datasets collected for southern California, are used to train the models and establish inference rules in order to predict the regional engineering parameters of the buildings in the region. The classification error and prediction power of the model are then presented in the paper and an example of the marginal probability distribution computed for a sample building is shown.

INTRODUCTION

Information extracted from remotely sensed data while creating 3-D models of urban areas is usually limited to spatial, spectral or geometric attributes of buildings. Spatial information includes, but is not limited to, attributes such as location of structures, i.e. longitude and latitude, proximity of structures and topography of the region. Spectral information contains attributes such as rooftop material, cladding or façade material, and etc. Geometric attributes provide information about height of structures, footprint area and perimeter of structures, degree of irregularity in plan view and elevation, dominant orientation of buildings in city blocks, roof types (flat, gable, hip and etc.) and other indices which can be derived form the geometry of objects. There is however another set of attributes, important in assessing vulnerability of structures subjected to natural or man-made disasters, which cannot directly be derived from remotely sensed data. Structural type, occupancy type and age of structures are among those attributes. Structural type is determined by the load resisting system (vertical and lateral) used in a structure. Examples of structural type are classes such as wood, steel, concrete and masonry structures. Occupancy type is defined as the social-use or the utility-class of structures. Examples of occupancy type include classes such as residential, commercial and industrial.

In this paper, a methodology for inferring structural type and occupancy type of buildings from other signatures and attributes of an urban area such as the ones that can

be derived from imagery is formulated. Since the response variables to be modeled, i.e. structural type and occupancy type, as well as some of the independent variables such irregularity of buildings or roof type of structures are categorical data, the statistical model to be used for inferencing should incorporate a categorical data mining framework. An overview of categorical data modeling is first presented in this paper. The introduced methodology is then applied to different datasets to illustrate the application of statistical pattern recognition techniques in structural attribute modeling.

METHODOLOGY

In order to perform an inclusive vulnerability assessment of an urban area, subjected to different hazard scenarios (natural or man-made), a comprehensive inventory of structures at risk is needed. Attributes included in such building inventories usually consist of a mixture of classes. There are attributes with a well-defined measurement scale such as height, square footage, perimeter and age of structures that are categorized as *quantitative variables*. There are also attributes such as structural type or occupancy type of buildings which do not have any natural measurement scale and are defined based on a set of levels or classes. These types of variables are known as *qualitative* or *categorical variables*. The input variables to a statistical model, i.e. observations, are known as *explanatory variables*, *independent variables* or *predictors*. These measurements are regarded as non-random measurements. The output(s) of statistical models are referred to as *response variables* or *dependent variables*. These are the outcome of models for a given set of observations (response variables) and are regarded as random variables which are free to vary in response to explanatory variables.

In order to infer the structural- and occupancy-type of buildings from geometric and spatial attributes of the built-environment, a statistical framework which incorporates both quantitative and qualitative variables should be utilized. In this paper, application of *multinomial logistic regression models* in inferring categorical attributes of urban areas is investigated. Models developed in this paper are then used to establish a set of inference rules using training datasets.

In *multinomial logistic regression models*, probability of a response variable, from the i^{th} observation, falling into the k^{th} category given a set of explanatory variables can be expressed by a multinomial probability distribution as shown below:

$$P(y_i = k \mid \boldsymbol{x_i}) = \pi_k(\boldsymbol{x_i}) = \pi_{ik} \quad i = 1:N \ \text{and} \ k = 1:K-1 \tag{4}$$

where ($\boldsymbol{x_i}$, y_i) is the i^{th} observation such that $\boldsymbol{x_i}$ = (x_{i1}, x_{i2}, …, x_{ip}) is a vector of p explanatory variables and y_i is the corresponding response variable. π_{ik} is the probability of i^{th} response variable falling in the k^{th} category.

It can be seen that for the i^{th} observation, the response variable with K categories can be treated as a multinomial variable with probabilities $\{\pi_{i1}, \pi_{i2}, ..., \pi_{iK}\}$ and the constraint that $\sum_{k=1}^{K} \pi_{ik} = 1$. To impose the constraint that fitted probabilities on the K categories

should sum to one, one of the categories should arbitrary be selected as the *base category* or the *control group*. This category can be the first, the last or any other. Choosing the last category as the baseline category, the *log-odds* or "logarithm of the ratio between logit model of the k^{th} category and the baseline category" for p explanatory variables can be expressed as shown in Equation 5.

$$\ln(\frac{\pi_{ik}}{\pi_{iK}}) = \sum_j \beta_{jk} x_{ij} = \alpha_k + \beta_{1k} x_{i1} + ... + \beta_{pk} x_{ip} \quad (5)$$

where α_k and β_{jk}'s are the logistic regression coefficients of the log-odds of the k^{th} category relative to the base category.

Using Equation 5, probability of i^{th} observation falling in the k^{th} category can then be expressed as below:

$$\pi_{ik} = \frac{\exp(\sum_j \beta_{jk} x_{ij})}{1 + \sum_{k=1}^{K-1} \exp(\sum_j \beta_{jk} x_{ij})} \; ; \; i = 1:N \; , \; j = i:p \quad (6)$$

and therefore $\pi_{iK} = 1 - \sum_{k=1}^{K-1} \pi_{ik}$ (7)

where π_{ik} is the probability of i^{th} observation falling in the k^{th} category.

Parameter estimation for log-odds of explanatory variables is done by maximizing the expectation of log-likelihood of each variable. For each log-odds, estimated parameters ($\hat{\boldsymbol{\beta}}$) as well as their standard error ($\sigma_{\hat{\beta}}$) can be calculated. The significance of each parameter in the model can be assessed using the *Wlad statistics*. The *Z-value* of Wald statistics for each parameter can be calculated by computing the ratio of estimated parameters and their standard error term as shown in Equation 8.

$$Z = \frac{\hat{\beta}}{\sigma_{\hat{\beta}}} \sim N(0,1) \quad (8)$$

where Z is the Wald statistics of the estimated parameter $\hat{\beta}$. The standard error of parameter $\hat{\beta}$ is shown by $\sigma_{\hat{\beta}}$.

The computed *Z-value* has a normal distribution and can be used to judge the significance of the coefficient. It can be shown that for large sample sizes, Z^2 has a chi-square distribution with one degree of freedom. To judge the overall suitability and parsimony of a model, the *Akaike Information Criterion* (*AIC*) is used. In the context of logistic models presented in this paper, the *AIC* can be defined as the sum of residual deviance of the model and the number of regression coefficients as shown in Equation 9.

$$AIC = -2L(\hat{\ }) + 2n \quad (9)$$

where $L(\hat{\ })$ is the maximum log-likelihood of the fitted model and *n* denotes total number of variables in the model.

Smaller values of *AIC* indicate a better fit to the data. *AIC* can also be used as a comparison tool when it comes to model selection.

DATASETS AND MODEL DEVELOPMENT

In this section, application of multinomial logistic regression model -explained in earlier parts of this paper- are discussed. Two sets of data from southern California, USA are collected and used in this study. The first set of data, also referred to as *dataset A* in this paper, is the aggregated tax assessor database of five counties in southern California originally developed for *EPEDAT* [1]. This database contains structural and occupancy attributes of buildings as well as height, total square footage and year of construction of structures at the "census tract" level. The database contains total of 38,135 buildings from 1,570 census tracts from counties of Los Angeles, Orange, Riverside, San Bernardino and Ventura as well as aggregated inventory data of the city of Los Angeles, which was originally excluded from the Los Angeles county dataset. The second set of data used in this paper, also referred to as *dataset B*, is the detailed inventory data of eighteen census tracts within the Orange county. The inventory database of Orange county is collected at the building level with attributes extracted from tax assessor database of the county as well as from remotely sensed imagery. In this dataset, structural type, occupancy type and year of construction of buildings are extracted from tax assessor databases while height, square footage, configuration in plan view as well as rooftop type of 1,947 buildings, are extracted from optical imagery.

In order to identify all the possible models that can be created using *dataset A*, the most primitive combination of attributes in this dataset, also known as the *baseline model*, should first be identified. Additional explanatory variables then will be added to the baseline model one at a time. The most basic model to be considered using *dataset A* is the one associated with only two explanatory variables; i.e. the height and square footage of buildings. This choice of variables is mainly because of the fact that height and footprint area of structures are the only two common attributes between *dataset A* and the ones that can directly be extracted from remotely sensed data when reconstructing 3-D city models. If additional information such as rooftop type, cladding, age of the buildings and etc. is available from auxiliary sources, it can be added to the model in the later stages. Finally, if either structural type or occupancy type of a building is known it is reasonable to include that attribute in the model in order to predict the other one. Therefore, for each of the response variables, i.e. structural type and occupancy type, three models can be created. Table 1 summarizes different models which can be created using *dataset A*, starting with a baseline model in each case. Supplementary explanatory variables are then added to the baseline model one at a time.

Table 1 Summary of models created from *dataset A*

Model ID	Response Variable	Explanatory Variable Included in the Model			
		Variable #1	Variable #2	Variable #3	Variable #4
Model I*	General Str. Type	Height (ft)	Ave. Area (ft^2)	-	-
Model II	General Str. Type	Height (ft)	Ave. Area (ft^2)	General Occ. Type	-
Model III	General Str. Type	Height (ft)	Ave. Area (ft^2)	General Occ. Type	Age
Model IV*	General Occ. Type	Height (ft)	Ave. Area (ft^2)	-	-
Model V	General Occ. Type	Height (ft)	Ave. Area (ft^2)	General Str. Type	-
Model VI	General Occ. Type	Height (ft)	Ave. Area (ft^2)	General Str. Type	Age

* Baseline model

In case of *dataset B*, both tax assessor files and remotely sensed imagery are used to compile the dataset, therefore this dataset has larger number of attributes associated with each building compare to *dataset A.* There are seven attributes associated with each observation. These are height, square footage, irregularity, rooftop type, year of construction, structural type and occupancy type. Following the argument made for *dataset A* regarding creating a baseline model upon the most primitive attributes which can directly be extracted from remotely sensed data and then adding more attributes to the model, it is reasonable to include the following attributes in the prime model of *dataset B*: height, square footage, irregularity and rooftop type. Depending on the availability of ancillary data, year of construction or age of buildings, occupancy type or structural type can be included in the model in the later steps. Table 2 summarizes different models which can be created using *dataset B*

Table 2 Summary of models created from *dataset B*

Model ID	Response Variable	Explanatory Variable Included in the Model					
		Variable #1	Variable #2	Variable #3	Variable #4	Variable #5	Variable #6
Model I*	Str. Type	Height (ft)	Area (ft^2)	Configuration	-	-	-
Model II	Str. Type	Height (ft)	Area (ft^2)	Configuration	Roof	-	-
Model III	Str. Type	Height (ft)	Area (ft^2)	Configuration	Roof	Occ. Type	-
Model IV	Str. Type	Height (ft)	Area (ft^2)	Configuration	Roof	Occ. Type	Age
Model V*	Occ. Type	Height (ft)	Area (ft^2)	Configuration	-	-	-
Model VI	Occ. Type	Height (ft)	Area (ft^2)	Configuration	Roof	-	-
Model VII	Occ. Type	Height (ft)	Area (ft^2)	Configuration	Roof	Str. Type	-
Model VIII	Occ. Type	Height (ft)	Area (ft^2)	Configuration	Roof	Str. Type	Age

* Baseline Model

Datasets *A* and *B* are used to compute the parameters of multinomial logistic models, fitted to each set of variables defined in Tables 1 and 2. In order to calculate the overall classification error of a multinomial logistic model, the corresponding classification table for the response variable using prediction rules defined by the model should be calculated. The diagonal elements of this table represent the number of correctly classified observations. Classification error can then be calculated by computing the ratio between sum of the diagonal elements of the table and total number of elements in the table as shown in Equation 10.

$$\varepsilon = 1 - \frac{sum[diag(T)]}{sum[T]} \tag{10}$$

where ε is the classification error or misclassification rate, in table *T* and *diag(.)* refers to the diagonal elements of the table.

Examples of classification tables for *Model I* through *Model III* of *dataset A* (in Table 1) are shown in Tables 3 through 5, respectively. A summary of the *AIC*, the degrees of freedom *df*, for estimating parameters of each model and the overall classification error of each of the models is presented in Tables 6 and 7, respectively.

Table 3 Classification table for structural classes using *Mode I* of Table1

		Predicted Structural Classes					
		C	C/S	RM	S	URM	W
Observed Structural Classes	C	**1590**	0	0	128	0	7575
	C/S	45	**0**	0	1	0	440
	RM	349	0	**0**	13	0	2842
	S	542	0	0	**149**	0	437
	URM	325	0	0	5	**0**	3151
	W	685	0	0	6	0	**16505**

Table 4 Classification table for structural classes using *Mode II* of Table1

		Predicted Structural Classes					
		C	C/S	RM	S	URM	W
Observed Structural Classes	C	**3917**	0	33	145	0	5198
	C/S	151	**0**	0	0	0	335
	RM	380	0	**48**	38	0	2738
	S	490	0	42	**186**	0	410
	URM	869	0	16	6	**0**	2590
	W	1333	0	49	41	0	**15773**

Table 5 Classification table for structural classes using *Mode III* of Table1

		Predicted Structural Classes					
		C	C/S	RM	S	URM	W
Observed Structural Classes	C	**6652**	0	60	164	12	2405
	C/S	401	**0**	7	0	0	78
	RM	715	0	**142**	0	0	2347
	S	704	0	0	**424**	0	0
	URM	1710	0	43	0	**0**	1728
	W	2484	0	122	0	0	**14590**

Table 6 Summary of *AIC*, degrees of freedom (*df*) and the overall classification error of models in Table 1 used in building the multinomial logistic regression models from *dataset A*.

Model ID	Response Variable	Explanatory Variable				*AIC*	*df*	Classification Error
		#1	#2	#3	#4			
Model I*	Str. Type	Height (ft)	Area (ft^2)	-	-	87,133	20	47.56%
Model II	Str. Type	Height (ft)	Area (ft^2)	Occ. Type	-	77,450	50	42.73%
Model III	Str. Type	Height (ft)	Area (ft^2)	Occ. Type	Age	58,285	60	37.31%
Model IV*	Occ. Type	Height (ft)	Area (ft^2)	-	-	91,825	24	51.14%
Model V	Occ. Type	Height (ft)	Area (ft^2)	Str. Type	-	81,894	54	46.44%
Model VI	Occ. Type	Height (ft)	Area (ft^2)	Str. Type	Age	81,608	66	46.69%

* Baseline model

Table 7 Summary of *AIC*, degrees of freedom (*df*) and the overall classification error of models used in building the multinomial logistic regression models from *dataset B.*

Model ID	Response Variable	Explanatory Variable						*AIC*	*df*	Classification Error
		#1	#2	#3	#4	#5	#6			
Model I*	Str. Type	H	A	Config.	-	-	-	3631	20	30.71%
Model II	Str. Type	H	A	Config.	Roof	-	-	2663	35	21.98%
Model III	Str. Type	H	A	Config.	Roof	Occ.	-	1827	50	17.36%
Model IV	Str. Type	H	A	Config.	Roof	Occ.	Age	1771	55	16.18%
Model V*	Occ. Type	H	A	Config.	-	-	-	3697	12	43.50%
Model VI	Occ. Type	H	A	Config.	Roof	-	-	2984	21	29.74%
Model VII	Occ. Type	H	A	Config.	Roof	Str.	-	2062	36	18.90%
Model VIII	Occ. Type	H	A	Config.	Roof	Str.	Age	2064	39	18.90%

* Baseline model

RESULTS AND EXAMPLES

The multinomial logistic model in each case can be used for classification purposes by defining a set of decision rules. For instance, for an input *attribute vector* -consisted of p independent values- to the model, i.e. $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$, probability of the response falling into each of the K categories can be computed as $\Pi = \{\pi_1, \pi_2, ..., \pi_K\}$ using Equations 6 and 7. The category corresponding to the highest probability in Π can be selected as the class to which the input attribute vector belongs. Furthermore, a minimum probability threshold can be chosen such that if the highest probability in Π is below that threshold, the classification results in an "unclassified" status. In cases in which the probability difference between two classes is not significant, a tie assignment between class-membership will result and therefore, a set of rules to assign the response variable to the correct class should be defined. It should be noted that in many cases decision rules depend on the nature of the input variables to the mode as well the resulted response variable, and hence they differ from one problem to the other. Therefore, careful consideration should be given while compiling the decision rules for a specific problem.

Table 8 shows the result of parameter estimation for *Model I* from *dataset A*. The *base-category* (also known as *control group*) for explanatory variables in this table is the structural class type "*C*", i.e. the concrete class, and is highlighted by an asterisk.

Table 8 Log-odds parameters of *Model I* from *dataset A* (Table 1)

Log-odds		Intercept	H* (High)	H (Medium)	H (Low)	Average Area
C/S	Coefficients	**-2.216**	-	**-0.324**	**-0.572**	**-1.3E-05**
	Std. Error	5.2E-11	-	5.3E-12	4.3E-11	2.2E-06
RM	Coefficients	**-0.574**	-	**0.337**	**-0.355**	**-1.2E-05**
	Std. Error	1.7E-11	-	2.7E-12	1.3E-11	8.3E-07
S	Coefficients	**-1.154**	-	**-0.130**	**-1.502**	**8.4E-06**
	Std. Error	5.3E-12	-	2.2E-12	2.6E-12	4.1E-07
URM	Coefficients	**-0.401**	-	**0.382**	**-0.388**	**-1.7E-05**
	Std. Error	2.5E-11	-	3.5E-12	2.0E-11	9.7E-07
W	Coefficients	**-8.096**	-	**11.408**	**9.099**	**-4.9E-05**
	Std. Error	3.6E-11	-	5.9E-12	3.1E-11	8.7E-07

* Base-category Variable

In order to assign a class to an independent observation, the probability vector Π should first be computed. The category corresponding to the highest probability in Π is then assigned to the observation. As an example let's assume the structural type of a low-rise building with average square footage of 2,176 ft^2 in southern California is to be predicted. Using the estimated parameters for the corresponding model, i.e. *Model I*, shown in Table 8 and using structural class *C*, i.e. concrete, as the base-category for the response variable, the log-odds ratios can be calculated as shown in Equation 11.

$$\begin{cases} \ln(\frac{\pi_{C/S}}{\pi_C}) = -2.216 - 0.324x_1 - 0.572x_2 - 0.000013x_3 \\ \ln(\frac{\pi_{RM}}{\pi_C}) = -0.574 + 0.337x_1 - 0.355x_2 - 0.000012x_3 \\ \ln(\frac{\pi_S}{\pi_C}) = -1.154 - 0.130x_1 - 1.502x_2 + 0.0000084x_3 \\ \ln(\frac{\pi_{URM}}{\pi_C}) = -0.401 + 0.382x_1 - 0.388x_2 - 0.000017x_3 \\ \ln(\frac{\pi_W}{\pi_C}) = -8.906 + 11408x_1 + 9.099x_2 - 0.000049x_3 \end{cases} \quad (11)$$

where x_1 and x_2 are dichotomous dummy variables corresponding to height of the structure as defined in Table 9. x_3, is a quantitative variable corresponding to average square footage of the structure in ft^2.

Table 9 Dummy variables x_1 and x_2 as indicators of height in Equation 11

Height (High-rise)	0	0
Height (Medium-rise)	1	0
Height (Low-rise)	0	1

Therefore, the marginal mass-probability density of the observation falling into different response categories can be computed, using $x_1 = 0$, $x_2 = 1$ and $x_3 = 2176$, as shown below:

$$\begin{cases} \pi_C = \dfrac{1}{1+e^{-2.817401}+e^{-0.9549136}+e^{-2.63715737}+e^{-0.8268078}+e^{+0.8956142}} = 0.2271515 \\ \pi_{C/S} = \dfrac{e^{-2.817401}}{1+e^{-2.817401}+e^{-0.9549136}+e^{-2.63715737}+e^{-0.8268078}+e^{+0.8956142}} = 0.01357467 \\ \pi_{RM} = \dfrac{e^{-0.9549136}}{1+e^{-2.817401}+e^{-0.9549136}+e^{-2.63715737}+e^{-0.8268078}+e^{+0.8956142}} = 0.0874174 \\ \pi_S = \dfrac{e^{-2.63715737}}{1+e^{-2.817401}+e^{-0.9549136}+e^{-2.63715737}+e^{-0.8268078}+e^{+0.8956142}} = 0.01625607 \\ \pi_{URM} = \dfrac{e^{-0.8268078}}{1+e^{-2.817401}+e^{-0.9549136}+e^{-2.63715737}+e^{-0.8268078}+e^{+0.8956142}} = 0.09936458 \\ \pi_W = \dfrac{e^{+0.8956142}}{1+e^{-2.817401}+e^{-0.9549136}+e^{-2.63715737}+e^{-0.8268078}+e^{+0.8956142}} = 0.5562357 \end{cases} \quad (12)$$

Hence, the probability vector $\Pi = \{\pi_C, \pi_{C/S}, \pi_{RM}, \pi_S, \pi_{URM}, \pi_W\}$ can be assembled as:

$$\Pi = \{0.2272\ ,\ 0.01357\ ,\ 0.0874\ ,\ 0.0163\ ,\ 0.0994\ ,\ 0.5562\}$$

It can be seen that the last category, i.e. *W*, in the probability vector Π has the highest value and therefore, in absence of any other information such as age or occupancy type, the predicted structural type of a low-rise building with an average square footage of 2,176 ft^2 is *wood frame*. This class prediction is in agreement with one of the observations from the *dataset A.*

CONCLUSIONS

The multinomial logistic models presented in this paper provide the means to compute quantitative measures of marginal building class-membership probabilities for categorical attributes associated with structures. Based on the comparison of similar models presented in Tables 6 and 7, it can be seen that the more detailed training database, i.e. *dataset B*, results in models with a better prediction rate. Furthermore, it can be seen that for the multinomial logistic models with similar explanatory variables, inferring the structural type generally results in a lower prediction error than inferring the occupancy type. Therefore, it can be concluded that in predicting marginal probabilities, use of structural type as the independent variable generally results in a smaller classification error than the occupancy type.

ACKNOWLEDGEMENT

This research was supported by the National Science Foundation (NSF) Grant EEC-9701568, the UPS Foundation Grant from Stanford University, the Multidisciplinary Center for Earthquake Engineering Research Grant EEC-9701471, and ImageCat Inc**.** The authors would like to gratefully acknowledge their supports.

The material presented in this paper is part of the doctoral dissertation of the first author [2]. Any opinions, findings and conclusions or recommendations expressed in this publication are those of the authors, and do not necessarily reflect those of NSF, MCEER, Stanford University or ImageCat Inc.

REFERENCES

[1] EPEDAT, 1994, "*Status Report: Early Post-Earthquake Damage Assessment Tool for Southern California*," Prepared by EQE International. November 1994.

[2] Sarabandi P., 2007, "*Development of Algorithms for Building Inventory Compilation through Remote Sensing and Statistical Inferencing,*" Ph.D. Dissertation, Department of Civil and Environmental Engineering, Stanford University, Stanford, CA.