On the selection of significant variables in a model for the deteriorating process of facades

C Serrat[1,5], **V Gibert**[2], **J R Casas**[3] **and J Rapinski**[4]

[1] Dept of Mathematics, IEMAE-EPSEB, Universitat Politècnica de Catalunya-BarcelonaTECH, Dr Marañón, 44-50, 08028-Barcelona, Spain
[2] Dept of Architectural Technology, LABEDI-EPSEB, Universitat Politècnica de Catalunya-BarcelonaTECH, Dr Marañón, 44-50, 08028-Barcelona, Spain
[3] Dept of Civil and Environmental Engineering, Universitat Politècnica de Catalunya-BarcelonaTECH, Jordi Girona, 1-3, 08034-Barcelona, Spain
[4] Institute of Geodesy, University of Warmia and Mazury, Oczapowskiego, 2, 10719-Olsztyn, Poland

E-mail: carles.serrat@upc.edu

Abstract. In previous works the authors of this paper have introduced a predictive system that uses survival analysis techniques for the study of time-to-failure in the facades of a building stock. The approach is population based, in order to obtain information on the evolution of the stock across time, and to help the manager in the decision making process on global maintenance strategies. For the decision making it is crutial to determine those covariates -like materials, morphology and characteristics of the facade, orientation or environmental conditions- that play a significative role in the progression of different failures. The proposed platform also incorporates an open source GIS plugin that includes survival and test moduli that allow the investigator to model the time until a lesion taking into account the variables collected during the inspection process. The aim of this paper is double: a) to shortly introduce the predictive system, as well as the inspection and the analysis methodologies and b) to introduce and illustrate the modeling strategy for the deteriorating process of an urban front. The illustration will be focused on the city of L'Hospitalet de Llobregat (Barcelona, Spain) in which more than 14,000 facades have been inspected and analyzed.

1. Introduction

In order to design a multiscale predictive platform for the analysis of the deterioration process of the facades in a building stock, we will focus on determining a reliable inspection method that fits to any urban laboratory. The system proposes a network of urban laboratories for building research analysis and information (BRAIN) with the aim of modeling the degradation time based on the information from all the cities participating in the project, from a multiscale perspective.

On the other hand, the methodology must study the morphology of the architectural elements and the states of the damages in the facades, from a general perspective. Basic fundamentals and

---

[5] To whom any correspondence should be addressed.

definitions on the methodological parts and preliminary works have been previously introduced by Serrat and Gibert [1] and Gibert et al. [2].

2. A predictive system

In this section we will introduce, in a short manner, the four ingredients of the predictive system that we are proposing. That is a) the collaborative approach in order to joint and analyze the information from the nodes in the network of urban labs, b) the inspection methodology to be applied to the urban lab, c) the survival analysis methodology as a statistical technique for the durability and modeling estimation and, d) the GIS platform as a tool for managing the information and the analyses. Details of these components can be found in [3].

2.1. A collaborative perspective

The research focuses the prospecting campaign of the facades at a multiscale level. Indeed, it concentrates the interest within the concept of the urban laboratory that collects the envelope of the buildings and constitutes the urban front. This approach defines the urban canyon as the U-shape location where the facades are exposured and the time-to-event variables as well as the territorial and environmental covariates occur and determine the facades durability over time. Figure 1 shows the urban laboratory and the urban canyon [3].

The project is a translatable study to any city in the world in a networking manner. The net has a neuronal analysis center, the Collaboratory, as a coordination unit, able to store and analyze the information from the cells in order to test for similarities and differences accros the participants (cities, i.e. urban labs). Right-hand part of the figure 1 illustrates the network. The main goal is to joint efforts into the predictive knowledge of the deterioration of the urban fronts and facilitate the design of common standards as well as protocols of followup and intervention strategy.
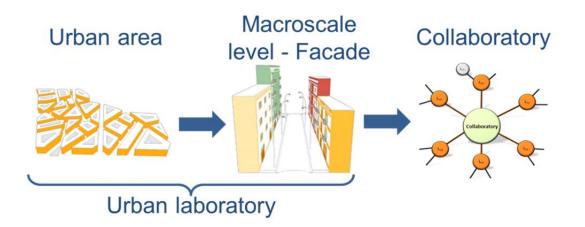

Figure 1. Urban laboratory network and Collaboratory for the BRAIN project.

2.2. Inspection methodology

The main goal of the monitoring part is the followup of the facades in order to detec in which particular moment of the service life of the facade the damages occur and progress. From this perspective Gibert et al [2] designed the inspection protocol based on a list of requirements in order to apply a population approach and as a result an inspection document was derived. The document consists of two parts. Part a) allows to collect field data, cartographic data, cadastral data as well as plot/building/facade data and architectural characteristics. Part b) is made for collecting the existing elements and materials and the state of damages at the time of inspection.

2.3. Survival analysis methodology

Let T be the time from the beginning of the follow-up (time zero) until the failure (the event of interest) happens. T is our random variable of interest. Let $f(t)$ and $F(t)$ be the density and the cumulative distribution function of the random variable T. Based on *f* and *F* we can derive statistics of interest like, the quantiles of the distribution. This will allow us to estimate the time until a proportion of damaged buildings in the population or, in the reverse sense, the proportion of damaged buildings at certain time for a particular damage. In the service life setup, time zero will mean the date that the building is built and by failure we will understand the successive grades of gravity, or the successive grades of extent, of the damages. The survival (i.e. durability) function for the random variable T is the complement to one of the distribution function F, that is,

$$S(t) = 1 - F(t) = 1 - \mathrm{P}(T \le t) = \mathrm{P}(T > t)$$

Our sample will be inspected at different inspection times (t_1, t_2, …) and the information on the failures times for the damages that the inspector collects will be interval censored, in the sense that the investigator only can ensure with probability one that the time to failure is in a time interval (l_i, r_i) [1].

In the context of our research on facades durability, since there are no references on the distribution functions that failure times follow, we will estimate the durability function and the hazard function in a non-parametric way. That is, our estimates will be only based on the data and we will not suppose any hypothetical (and non-testable) distributions family for the unknown density *f*. We will use the Turnbull's estimator [4] based on an iterative algorithm that maximizes the non-parametric likelihood function

$$L = \prod_{i \in O} \left(F(o_i) - F(o_i^-)\right) \prod_{i \in R} \left(1 - F(r_i)\right) \prod_{i \in L} F(l_i) \prod_{i \in I} \left(F(r_i) - F(l_i)\right),$$

where *O*, *R*, *L* and *I* are the subsets of exact, right-censored, left-censored and interval-censored observations, respectively. From the resulting probabilities, durability and hazard functions can be derived. From the computational point of view, in order to obtain the proper estimates for the survival model, we will use the Turnbull's estimator implemented in the statistical environment R [5] in the `Icens` package [6] available at Bioconductor website.

2.4. GIS Platform

In order to visualize and analyze, in a territorial and multiscalar manner, the available information on the urban canyon, we have implemented a QGIS [7] plugin for the Linux environment. The application allows the manager the visual followup (i.e. location and characteristics at a multiscalar level) of the ageing process of the urban laboratory. The language used to implement this application has been Python [8], open source language, object-oriented, which allows the use of other external programs, such as the case of R used for the statistical analysis. More details on the plugin and its development can be found in [3].

3. A modeling strategy

As a first result the proposed platform allows the investigator to select a subsample of facades of interest, a particular failure in a particular element and to estimate and draw the durability functions for each level of severity and each level of exten, in a marginal way. This output allows to compute the percentiles distribution as well as the cumulative distribution function of the overall subsample. In a similar manner, the plugin offers the possibility of adding one specific covariate value (material, orientation, enviromental condition…) and to derive the durability function for this specific value of that covariate. Based on this features it is possible to visually identify which variables could be significantly involved in the deterioration process.

In order to select the significant variables that take part in a non-parametric model for the deteriorating process we will use the extension of the Fleming-Harrington class of tests proposed for interval censored data by Oller and Langohr [9]. These tests allow to compare for equality of distribution functions and to check for ordered survival distributions among the categories of a covariate.
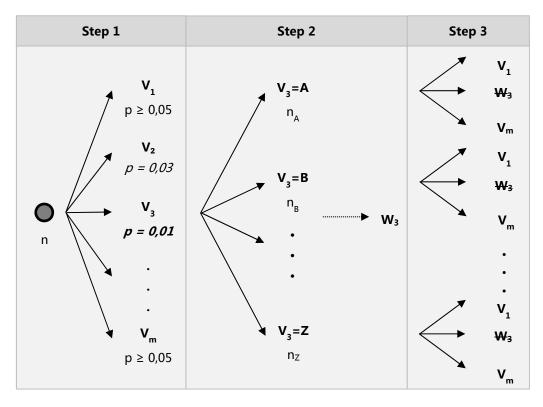
3.1. Proposed selection algorithm

The algorithm that we have designed to select, from a set of candidates, the sequence of covariates in the model for a given significance level, α, is:

STEP 1: Perform an equality test for each one the covariates candidates to explain the failure of interest. Choose, if exists, the covariate that shows the minimum p-value lower than α. In case none of the candidates satisfies the condition this would mean that none of the covariates would be significative enough to explane the deteriorating process.

STEP 2: Branch the subsample according the selected covariate and compute the sample size of each one the categories. Explore and validate the possibility of recoding those values. In case of recodification, apply equality tests in order to ensure the significancy of the just restratified covariate.

STEP 3: Repeat steps 1 and 2 for each one of the resulting branches by considering the remaining candidate covariates, until none of them becomes significative.

Figure 2 illustrates, in a global manner, the three steps of the proposed algorithm. The final non-parametric model will identify a significantly different behaviour in the aging process among the branches in the resulting classification tree.

Figure 2. Algorithm for the selection of the covariates of interest from a given set V_1, V_2, … V_m of candidates (for a supposed significance level $\alpha = 0.05$, i.e. a 95% of confidence level).

4. A case study: L'Hospitalet de Llobregat

The collaborative strategy has been introduced in a variety of countries and cities, as a strategy for building stock management [1]. As an illustration of the modeling strategy we will use the analysis of a particular failure in L'Hospitalet de Llobregat (hereafter L'Hospitalet) in the metropolitan area of Barcelona. L'Hospitalet is the city having the second largest population and is among the twenty most populated cities in Spain. The municipality covers an area of 12.5 square kilometres next to Barcelona, and it is divided in 12 neighborhoods.

A first group of 13,193 inspections were done by 2001 across all the city. Later on, by 2016, 1,308 facades have been reinspected. We will focus in the dwellings buildings located in the Centre, Collblanc and Santa Eulàlia neighborhoods as a statistical population of interest. The sample size is 814 and we are talking almost 150,000 squared meters of exposured facades. Figure 3 shows the subsample of interest.
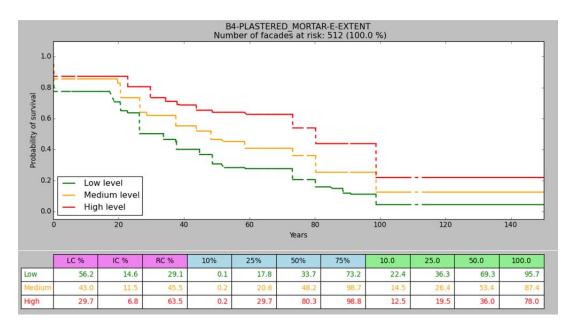

Figure 3. QGIS map of the subsample of interest in L'Hospitalet
de Llobregat (inspected facades in light blue, $n = 814$).

After computing the Weighted Severity Index [3] for all the failures of interest in the selected sample we saw that the most relevant failure is the moisture in the deck railing plasters. The subsample at risk of suffering this failure is made by 512 facades. Figure 4 shows, numerically and graphically, the estimated survival functions for each one of the level of extent (low, medium and high). We will focus now in obtaining a model non-parametric for the moisture in the deck railing plasters in a low level of extent (i.e. the dark green stepwise line in figure 4).
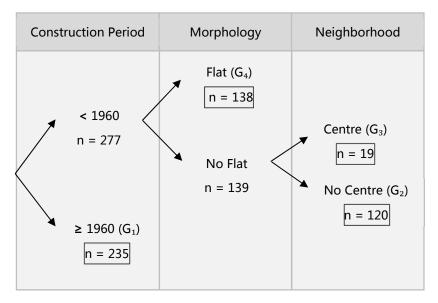
The covariates that have been considered of interest to be in the model are: V_1: the morphology of the facade, V_2: the construction period, V_3: the material, V_4: the neighborhood and V_5: the orientation. After applying the selection algorithm described in the previous section we obtain the model

$$T \sim V_2 + W_1 + W_4$$

where V_2 stands for the construction period stratified in two groups (before and after 1960), W_1 describe the morphology of the facada (flat versus no flat) and W_4 indicates if the facade belongs to the Center neighborhood. In figure 5 we can see the characteristics of the resulting four groups, named G_1, G_2, G_3 and G_4.

B4-PLASTERED_MORTAR-E-EXTENT
Number of facades at risk: 512 (100.0 %)

	LC %	IC %	RC %	10%	25%	50%	75%	10.0	25.0	50.0	100.0
Low	56.2	14.6	29.1	0.1	17.8	33.7	73.2	22.4	36.3	69.3	95.7
Medium	43.0	11.5	45.5	0.2	20.6	48.2	98.7	14.5	26.4	53.4	87.4
High	29.7	6.8	63.5	0.2	29.7	80.3	98.8	12.5	19.5	36.0	78.0

Figure 4. Durability functions for moisture in the deck railing plasters, in our subsample of interest in L'Hospitalet de Llobregat (% of censoring –in pink-, quantiles –in blue- and cumulative of failures across time –in green-).
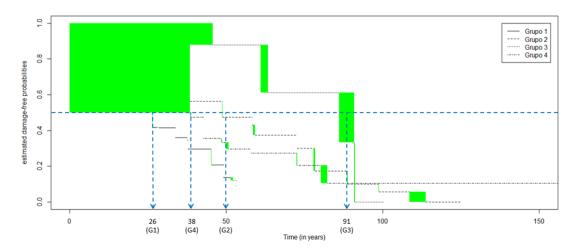

Figure 5. Classification tree for the time until the occurrence of puntual humidities in the plaster of the deck railing in the facades of residential buildings of L'Hospitalet de Llobregat.

By performing pairwise tests we can prove that the durability functions are ordered according to the groups in which the subsample has been splitted. In fact, the survival distributions follow the sequence

$$S_{G3} > S_{G2} > S_{G4} > S_{G1}$$

as it is shown in figure 6.

Figure 6. Stratified durability functions for puntual humidities in the plaster of the deck railing according the model defined by the construction period, the morphology and the neighborhood covariates, and estimated median time for each group.

It is interesting to highlight the improvement introduced in the predictive estimation from the proposed model with respect to the marginal estimation in figure 4. For example, the estimation of the median time to injury is approximately 34 years in the marginal model (figure 4), whereas, when the model is known, this median is estimated in about 26, 38, 50 and 91 years, for the facades of groups G_1, G_4, G_2 and G_3, respectively.

It can be observed that young facades (the ones in the group G_1) fails before than the others. In contrast, the facades with a better durability (the ones in the group G_3) are the older ones, located in the Centre neighborhood and with a no flat morphology (i.e. with other elements in the facade like balconies and tribunes). This fact poses a new and interesting debate on the construtive models and the associated building technology.

5. Conclusion

From a non-parametric perspective it is really convinient to model how different covariates can play a significant role in the aging process of the elements of the facades. When data are interval censored this goal is actually more challenging due to the uncertainty in the data. The main benefit of the proposed algorithm is that it allows to build a model in which at every step the most significative covariate is chosen and validated, in a sequential manner. This is a powerful strategy in ordre to identify priorities and to guide the decision making. Through the case study it has been illustrated the advantages of the modeling strategy by overpassing the limitations of not being available a parametric alternative.

References

[1]   Serrat C and Gibert V 2011 Survival analysis methodology for service live prediction and building maintenance, in *Proceeding of the 12th International Conference on Durability of Building Materials and Components*, vol. II, Porto, Portugal, 599-606.

[2]   Gibert V, Serrat C and Casas J R 2014 Determination of criteria for the exploration and for obtaining indicators in evolutional analysis of degradation in urban facades, in *Proceeding of the 13th International Conference on Durability of Building Materials and Components*, Sao Paulo, Brasil, 656-63.

[3]   Gibert V 2016 Sistema predictivo multiescala de la degradación del frente urbano edificado (in spanish), PhD Thesis, Universitat Politècnica de Catalunya-BarcelonaTECH.

[4]   Turnbull B W 1976 The empirical distribution function with arbitrarily grouped, censored and truncated data *Journal of the Royal Statistical Society, Series B* **38** pp 290-5.

[5]   R Core Team 2014 R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/.

[6]   Gentleman R and Vandal A 2016 Icens: NPMLE for Censored and Truncated Data, R package version 1.44.0.

[7]   QGIS Development Team 2016 QGIS Geographic Information System. Open Source Geospatial Foundation Project URL. http://www.qgis.org/

[8]   van Rossum G and the Python development team 2016 The Python Language Reference, release 2.7.12. URL: https://docs.python.org/2.7/

[9]   Oller R and Langohr K 2015 FHtest: Tests for Right and Interval-Censored Survival Data Based on the Fleming-Harrington Class, R package version 1.3.