

A Nonparametric Statistical Model for the Selection of Significant Variables Acting in the Deterioration of Built Façades

Carles Serrat¹, Vicenç Gibert², Joan R. Casas³ and Jacek Rapinski⁴

¹ IEMAE-EPSEB, Dept. of Mathematics, Universitat Politècnica de Catalunya-BarcelonaTECH, Av. Dr. Marañón, 44-50, 08028-Barcelona, Spain, carles.serrat@upc.edu

² LABEDI-EPSEB, Dept. of Architectural Technology, Universitat Politècnica de Catalunya-BarcelonaTECH, Av. Dr. Marañón, 44-50, 08028-Barcelona, Spain, vicenc.gibert@upc.edu

³ Dept. of Civil and Environmental Engineering, Universitat Politècnica de Catalunya-BarcelonaTECH, Jordi Girona, 1-3, 08034-Barcelona, Spain, joan.ramon.casas@upc.edu

⁴ Institute of Geodesy, University of Warmia and Mazury, Oczapowskiego, 2, 10719-Olsztyn, Poland, jacek.rapinski@uwm.edu.pl

Abstract. *Serrat and Gibert (2011) introduced the use of survival analysis techniques for the study of time-to-failure in the facades of a building stock. Those techniques have been implemented in a predictive system that incorporates the inspection methodology together with the statistical methodology. The system proposes a network of urban laboratories for building research analysis and information with the aim of modelling the degradation time based on the information from all the cities participating in the project, from a multiscale perspective. The approach is population based, that is from the manager of the building stock point of view, in order to obtain information on the evolution of the stock across time, and to help the manager with decision making process on global maintenance strategies. However, for an efficient decision making it is crucial to determine those covariates -like materials, morphology and characteristics of the facade, orientation or environmental conditions- that play a significant role in the progression of different failures along the façade service life. 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 damage when taking into account the variables collected during the inspection process. The aim of the paper is to introduce the methodology and the modelling strategy for the deteriorating process of an urban front. The contribution will be illustrated with a case study located in the city of L'Hospitalet de Llobregat (Barcelona, Spain) in which more than 14,000 facades have been inspected and analyzed.*

Keywords: *Non-Parametric Modelling, GIS for Building Durability, Inspection Methodologies, Survival Analysis, Collaboratory.*

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 definitions on the methodological parts and preliminary works have been previously introduced by Serrat and Gibert (2011) and Gibert *et al.* (2014).

The manuscript is organized as follows. In Section 2 the main predictive and methodological issues will be presented. Section 3 will be devoted to the non-parametric estimation of the statistical model and Section 4 will fully illustrate with a case study the Geographic Information System tool as well as the results and interpretation of the significant variables acting on the deterioration of built facades. The paper ends with a summary and the main conclusions of the conducted research.

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 join 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 Gibert (2016).

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 exposed 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 (Gibert, 2016).

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 join efforts into the predictive knowledge of the deterioration of the urban fronts and facilitate the design of common standards as well as protocols of follow-up and intervention strategy.

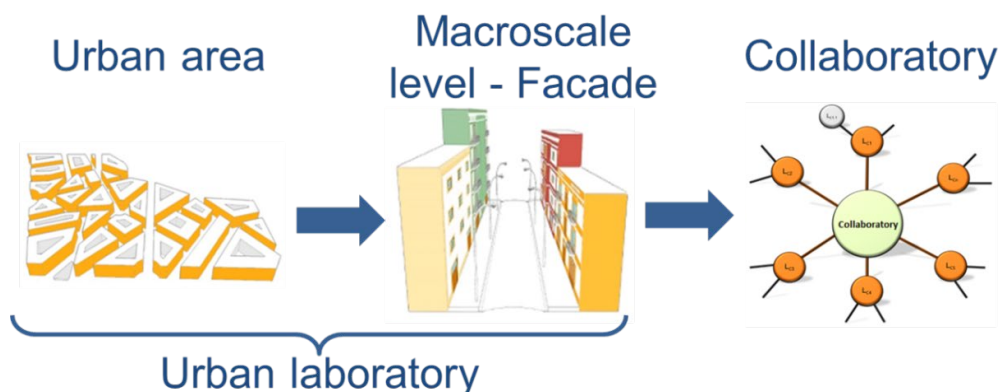


Figure 1. Urban laboratory network and Collaboratory for the BRAIN project.

2.2 Inspection Methodology

The main goal of the monitoring part is the follow-up of the facades in order to detect in which particular moment of the service life of the facade the damages occur and progress. From this perspective Gibert *et al.* (2014) 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,

$$S(t) = 1 - F(t) = 1 - P(T \leq t) = P(T > t) \quad (1)$$

Our sample will be inspected at different inspection times (t_1, t_2, \dots) 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) (Serrat and Gibert, 2011).

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 manner. 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 (Turnbull, 1976) based on an iterative algorithm that maximizes the non-parametric likelihood function

$$L = \prod_{i \in O} (F(o_i) - F(o_i^-)) \prod_{i \in R} (1 - F(r_i)) \prod_{i \in L} F(l_i) \prod_{i \in I} (F(r_i) - F(l_i)), \quad (2)$$

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 (R Core Team, 2014) in the Icen package (Gentleman and Vandal, 2016) available at Bioconductor website.

2.4 GIS Platform

In order to visualize and analyze, in a territorial and multiscale approach, the available information on the urban canyon, we have implemented a QGIS plugin for the Linux environment (QGIS Development Team, 2016). The application allows the manager the visual follow-up (*i.e.* location and characteristics at a multiscale level) of the ageing process of the urban laboratory. The language used to implement this application has been Python (van Rossum and the Python development team, 2016), 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 (Gibert, 2016).

3 A Modelling 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 extent, 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, environmental condition...) and to derive the durability function for this specific value of that covariate. Based on these 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 (2015). 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 constrain this would mean that none of the covariates would be significative enough to explain 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 significance 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 behavior in the aging process among the branches in the resulting classification tree.

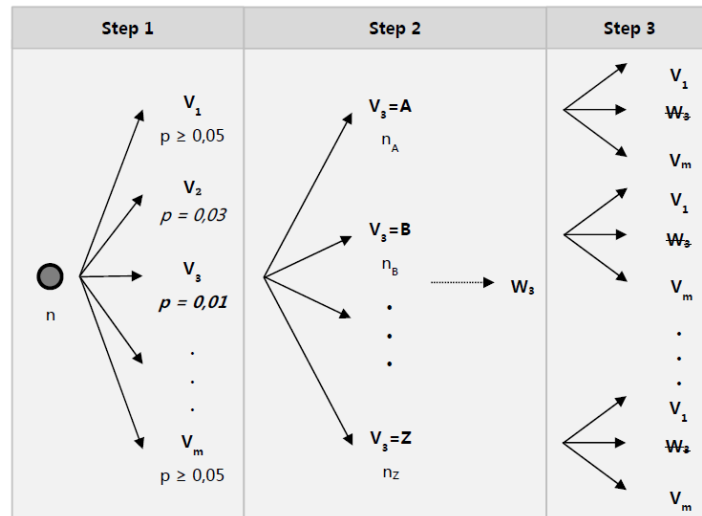


Figure 2. Algorithm for the selection of the covariates of interest from a given set V_1, V_2, \dots, V_m of candidates (for a supposed significance level $\alpha = 0.05$, *i.e.* a 95% of confidence level).

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 (Serrat and Gibert, 2011). As an illustration of the modelling 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 kilometers 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 dwelling 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 square meters of exposed facades. Figure 3 shows the subsample of interest.

After computing the Weighted Severity Index (Gibert, 2016) 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 levels 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

$$T \sim V_2 + W_1 + W_4 \quad (3)$$



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

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

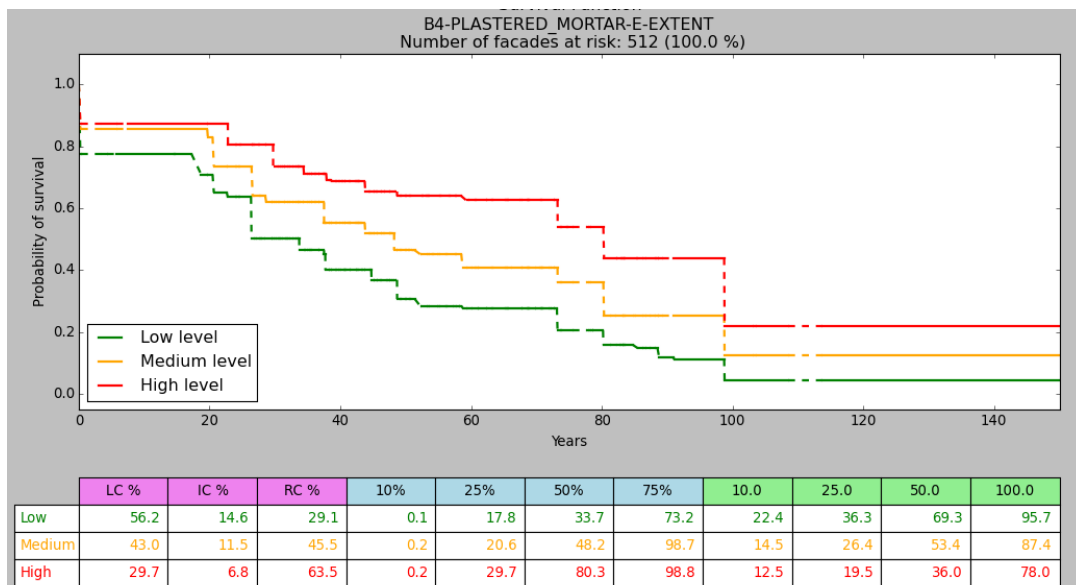


Figure 4. Durability functions for moisture in the deck railing plasters, in our subsample of interest in L'Hospitalet de Llobregat (% of censoring –in pink–, quantiles –in blue– and cumulative of failures across time –in green–).

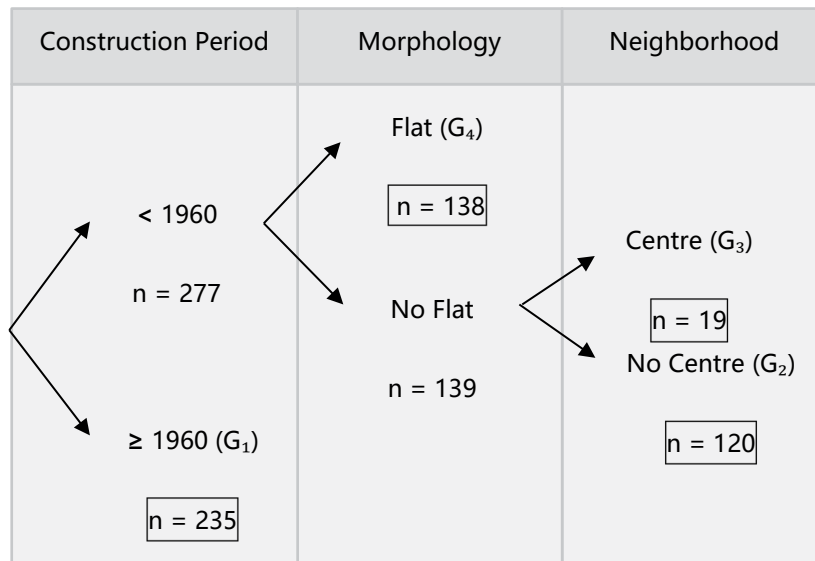


Figure 5. Classification tree for the time until the occurrence of puntual humidities in the plaster of the deck railing in the facades of residential buildings of L'Hospitalet de Llobregat.

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

$$S_{G_3} > S_{G_2} > S_{G_4} > S_{G_1} \tag{4}$$

as it is shown in Figure 6.

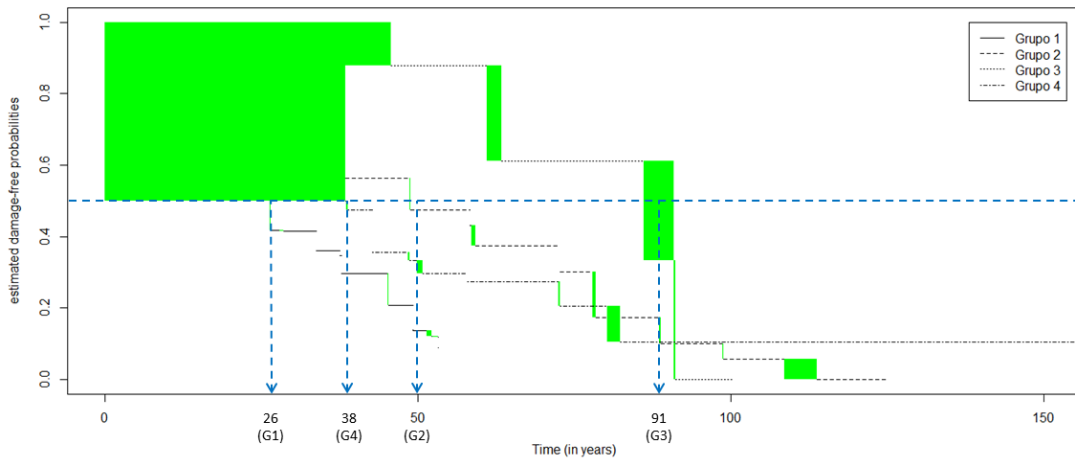


Figure 6. Stratified durability functions for puntual humidities in the plaster of the deck railing according the model defined by the construction period, the morphology and the neighborhood covariates, and estimated median time for each group.

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, 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 constructive models and the associated building technology.

5 Conclusions

From a non-parametric approach it is really convenient 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 significant covariate is chosen and validated, in a sequential manner. This is a powerful strategy in order to identify priorities and to guide the decision making. Through the case study it has been illustrated the advantages of the proposed modelling strategy overcoming the limitations of not being available a parametric alternative.

Acknowledgements

This research has been partially supported by grants MTM2015-64465-C2-1-R (MINECO / FEDER) from the Ministerio de Economía y Competitividad (Spain) and 2017 SGR 622 from the Departament d'Economia i Coneixement de la Generalitat de Catalunya. Authors are grateful to members of the IEMAE, LABEDI and GRASS-GRBIO groups their valuable comments and suggestions in the development of the work. Special thanks to Laura and Sandra for their implication and efficient collaboration in the project.

ORCID

Carles Serrat: <http://orcid.org/0000-0002-1504-5354>
 Vicenç Gibert: <http://orcid.org/0000-0001-6341-5762>
 Joan Ramon Casas: <http://orcid.org/0000-0003-4473-4308>
 Jacek Rapinski: <http://orcid.org/0000-0002-8954-7963>

References

- Gentleman, R. and Vandal, A. (2016). *Icens: NPMLE for Censored and Truncated Data*, R package version 1.44.0.
- Gibert, V. (2016). *Sistema predictivo multiescala de la degradación del frente urbano edificado* (in spanish), PhD Thesis, Universitat Politècnica de Catalunya-BarcelonaTECH.
- Gibert, V., Serrat, C. and Casas, J.R. (2014). Determination of criteria for the exploration and for obtaining indicators in evolutionary analysis of degradation in urban facades. In *Proceeding of the 13th International Conference on Durability of Building Materials and Components*, Sao Paulo, Brasil, 656–663.
- 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.
- QGIS Development Team (2016). *QGIS Geographic Information System*. Open Source Geospatial Foundation Project URL. <http://www.qgis.org/>
- 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/>
- 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.
- Turnbull, B.W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B*, 38, 290–295.
- van Rossum, G. and the Python development team (2016). *The Python Language Reference*, release 2.7.12. URL: <https://docs.python.org/2.7/>