The open discussion version of this paper is available at: Antoni A, Dhorne T. (2016)
Building size modelization.
PeerJ Preprints 4:e2264v3 https://doi.org/10.7287/peerj.preprints.2264v3

# Building size modelization

**Arlette Antoni**[1] **and Thierry Dhorne**[1]

[1]**Lab-STICC - CNRS UMR 6285, Université de Bretagne Sud, 56000 Vannes France**

## ABSTRACT

New challenges in the efficient management of cities depend on a deep knowledge of their inner structures. It is therefore very important to have access to reliable models of cities characteristics and organization. This paper aims at providing and validating a stochastic modelization based on statistical data of buildings parameters which can be useful as an entry for many other models considered in a wide range of fields where buildings structure is a main factor of a thorough modelization of cities. The interest of such an approach is highlighted through the detection of errors in the data or as a tool for visual clustering.

Keywords: Building size distributions, building geometry, statistical modelization, R language

## INTRODUCTION

As quoted by Michael Batty in 2008 ([2]) but still up to now " hardly any work exists on the properties of spatial distributions within individual cities" and particularly in what concerns buildings. A better knowledge of analytical properties of buildings is necessary in order to manage the key problems of to-day cities, in the sense that buildings support an important part of the general problems concerned by sustainable development and urban planning.

The usual way to manage building information in physical models is rather analytic and indeed not many work has been done on stochastics approach in this field.

The aim of the present talk is to provide elements of stochastic modelization of building characteristics with a special emphasize on size (area, height,...).

The level of modelization is the individual level, i.e. the building is considered in itself. Of course, a higher level, specially a spatial level should also be considered to cope with spatial organization structures, but this level is highly depending on the first level analysis coped with in this paper.

This type of modelization appears to be necessary (and relatively efficient) for :

- quality of initial data (corrections, normalizations,...),
- enrichment of data,
- classification (supervised of unsupervised) of city (districts),

the last point probably requiring a spatial modelization.

The study is concentrated on french urban areas but could be extended to other European countries by means of suitable data thanks to the french ANR project MAPUCE :
`http://www.cnrm-game-meteo.fr/spip.php?article787&lang=fr`.

Please cite this paper as: Arlette Antoni, Thierry Dhorne (2018) Building size modelization. In Marchesini I. & Pierleoni A. (Eds.)
Proceedings of the 4th Open Source Geospatial Research and Education Symposium (OGRS2016). Perugia, 12-14 October 2016.
https://doi.org/10.30437/ogrs2016_paper_32

## DATA

Though the general framework of this study concerns french urban areas, the data studied are restricted to some specific french cities. The initial data treatments are indeed rather heavy due to:

- the number of buildings,
- the necessity to control even roughly data quality,
- the need for an expert knowledge
- the specificity of the calculations.

Data come from french topographic data base: BD-TOPO. In this data base, each building is represented as a geometrical object and more precisely as a polygon (figure 1).

Classical spatial operators are then applied to these raw data in order to extract spatial attributes (or variables) from perimeter, area, up to convexity index and fractal dimension,...

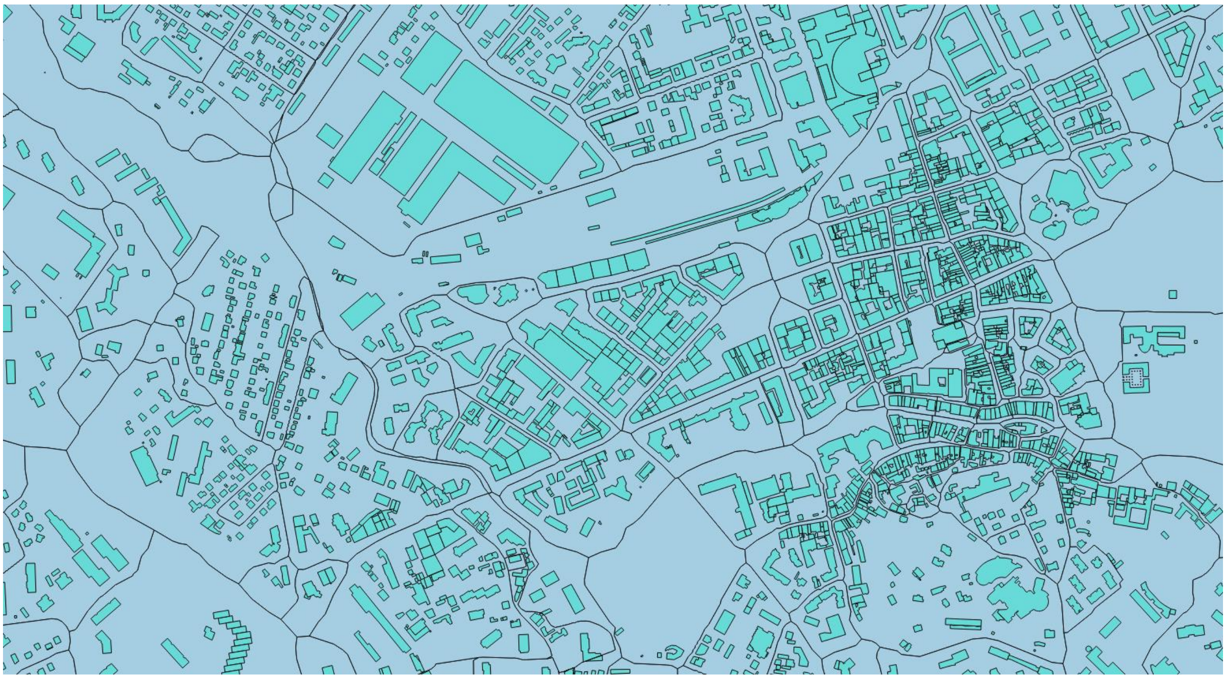

**Figure 1.** Representation of the GIS buildings layer

The study has been focused on five mid to big french cities:

| City | Area km2 | Inhabitants nr (2013) | Buildings nr |
|---|---:|---:|---:|
| Annecy | 13.65 | 52 029 | 37 271 |
| La Rochelle | 28.43 | 74 344 | 78 896 |
| Mulhouse | 22.18 | 112 063 | 126702 |
| Nantes | 65.19 | 292 718 | 262643 |
| Vannes | 32.30 | 53 032 | 16448 |

## DATA ANALYSIS

The whole data analytics is supported by the R language for statistical analysis. The area is the main variable studied. The initial data analysis reveals a high skew-symmetry of this variable when narrow binned histograms are considered.

Such skew-symmetries suggest the use of a logarithmic transformation which better reveals the essence of the phenomenon involved as represented in figure 2.

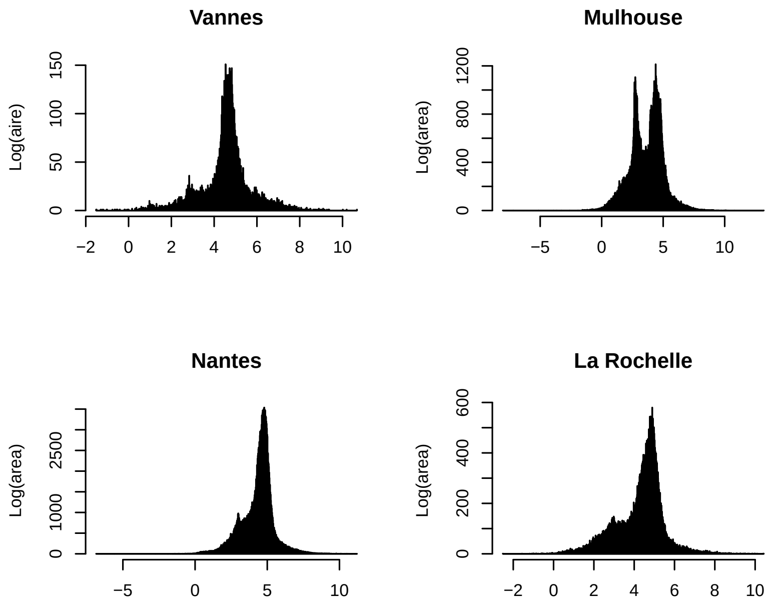

**Figure 2.** Histograms of log-transformed data

These histograms (figure 2) let appear a relative but meaningful multi-modality which can be important in some cities (for instance Mulhouse). The rather sharp distribution suggest that a gaussian model, usually considered by non specialists, is unsuited. Moreover, it seems that the slopes of left and right parts of the distributions are somewhat different. A thorough study of the data is then necessary.

## LITTERATURE MODELS

Size models have been widely used in the literature but as quoted before not for within cities studies and particularly for buildings studies.

First attempt to modelize size variables have been proposed for particles and then extended to other topics. The main models considered are the followings:

- Weibull or Rosin Rammler distribution is a useful distribution for representing particle size distributions generated by crushing operations. Its density is:

$$f(x,\lambda,k) = \frac{k}{\lambda}(\frac{x}{\lambda})^{k-1} \exp -(\frac{x}{\lambda})^k$$

with K, $\lambda$ shape and scale parameters.

- Log Normal distribution is often used to approximate the particle size distribution. Its density (on a logarithmic scale) is:

$$f(x) = \frac{1}{\sigma * \sqrt{2\pi}} \exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2)$$

- Log Hyperbolic distribution was proposed by Bagnold and Barndorff-Nielsen ([1]) to model the particle-size distribution of naturally occurring sediments.

$$\frac{1}{2\delta\sqrt{1+\pi^2 K_1(\zeta)}} \exp -\zeta\sqrt{1+\pi^2}\sqrt{1+(\frac{x-\mu}{\delta})^2} - \pi(\frac{x-\mu}{\delta}))$$

where $K_1()$ is the modified Bessel function of the third kind and order 1.

- Log skew Laplace model was proposed by Fieller, Gilbertson and Olbricht ([3]) as a simpler alternative to the log-hyperbolic distribution. Its density (on a logarithmic scale) is:

$$f(x) = \begin{cases} (\frac{1}{\alpha+\beta}) \ \exp(\frac{\mu-x}{\alpha}), & x \leq \mu \\ (\frac{1}{\alpha+\beta}) \ \exp(\frac{x-\mu}{\beta}), & x > \mu \end{cases}$$

with $\alpha$ and $\beta$ the left and right slopes, in the symmetric case $\alpha = \beta$

- Log double Pareto distribution appears also to be suitable as it is easier to manage than hyperbolic. Its density (on a logarithmic scale) is:

$$f(x) = \begin{cases} = \frac{\theta}{2\beta} \ (\frac{x}{\beta})^{\theta-1}, & 0 < x \leq \beta \\ \frac{\theta}{2\beta} \ (\frac{\beta}{x})^{\theta+1}, & x > \beta \end{cases}$$

Distributions functions are either power-law or exponential. The model complexity comes from the number of parameters. Four out of these five models are based on logarithmic transformation.

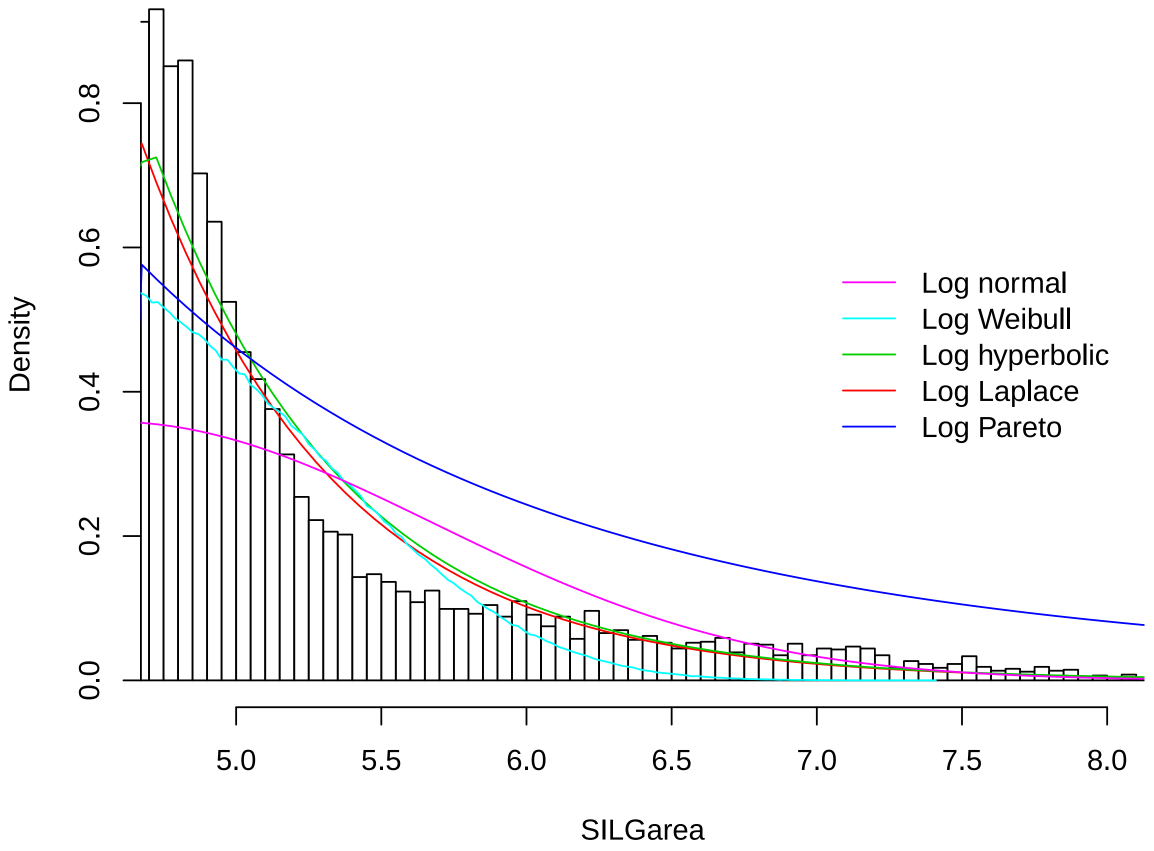

**Figure 3.** Comparison of fitted distributions

## COMPONENT MODELIZATION

The search for a suitable model for individual components identified in the population is rather difficult in presence of other mixed components. As seen on the histograms presented above, there is fortunately a major component corresponding to the bigger sizes. It is therefore possible to isolate such components in each city and to perform an analysis of the type of decreasing frequency distribution in order to choose the best model. Among modelization tested (see figure 3), Weibull and Log Normal appear to be poorly efficient. On the other hand, Log Hyperbolic is not suited for modelling tails. Attention is then restricted to the two remaining distributions which, though not optimal appear more relevant than the usual models.

## MIXTURE PARAMETER ESTIMATION

Once selected the appropriate model for any of the components, it is possible to modelize the data by a mixture of such components. The remaining problem consists in estimating the individual parameters of each component and the proportion (or conditional probability) of each (minus 1 due to the normalization constraint).

This has been done by maximum likelihood estimation for the Log Laplace case which is indeed easier to manage than Log hyperbolic case. An algorithm based on expectation-maximization has been developed and applied to the five cities data. A set of R functions has been developed to ease the use of these techniques. When applied to Vannes city, the optimal clustering let appear 6 classes segmented as in figure 4.

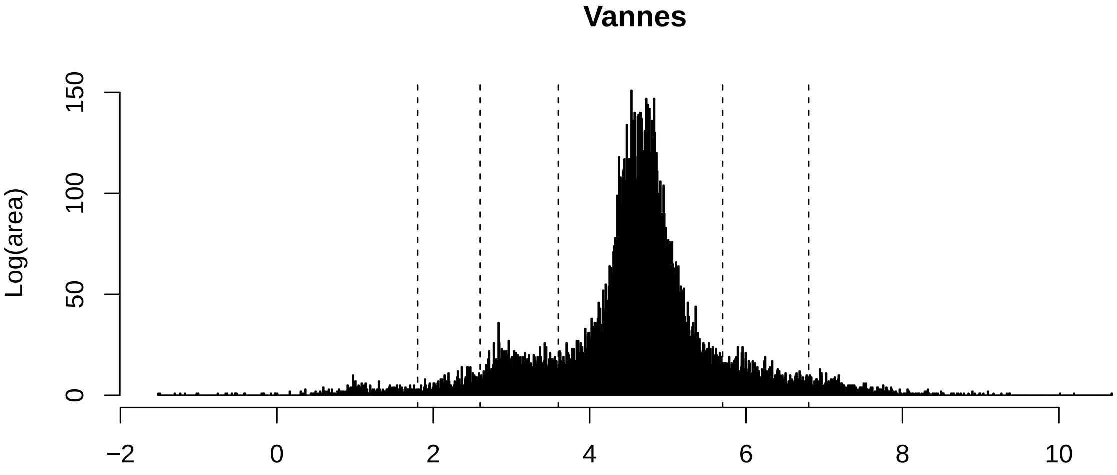

**Figure 4.** Histogram : log-transformed data of Vannes

## APPLICATION

The clusters identified can then be studied individually in order to visualize their locations in the city. In the case of one of the Vannes clusters, as shown in figure **??**, a strange aggregate of point is revealed. After thorough examination of the initial data, it appears that these points correspond to messy data.

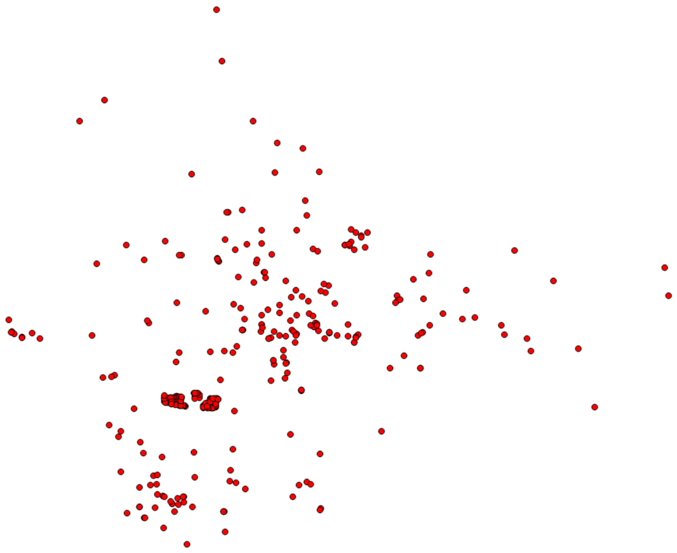

**Figure 5.** small size cluster location

The clustering procedure can also help to identify a spatial structure of cities. Still in the case of Vannes, a graphic of the different clusters (see figures 6 and 6) shows this type of structure and is therefore useful for a visual clustering of cities.

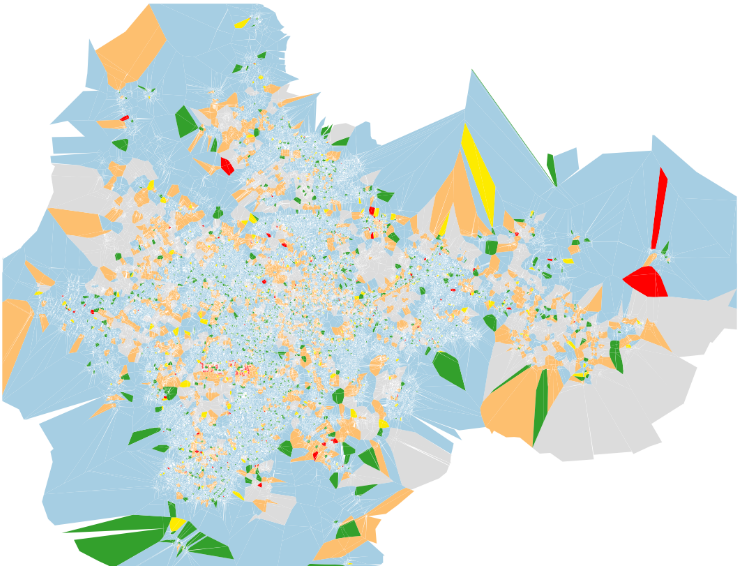

**Figure 6.** Representation of the different class

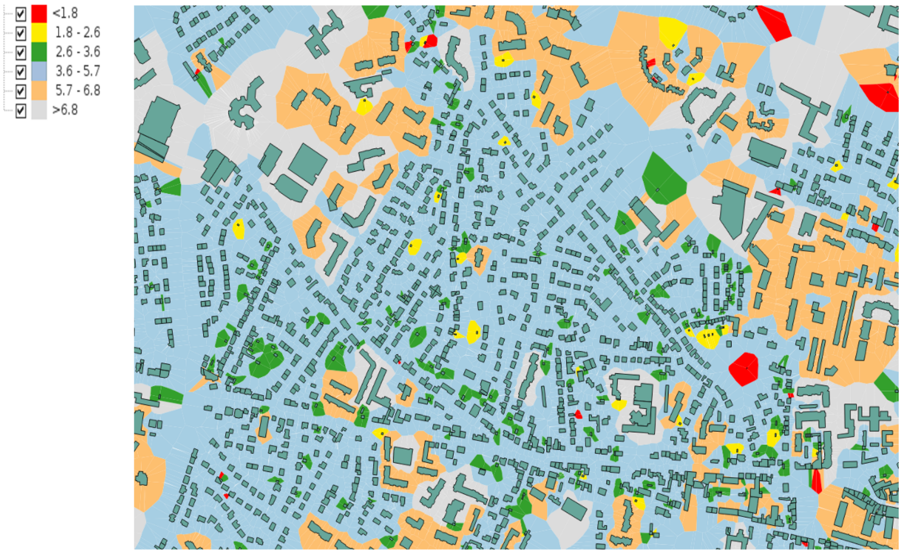

**Figure 7.** Zoom

## CONCLUSION

An initial data analysis of the buildings size variables let appear to majors problems that should be tackled when trying to create indicators for further analyis:

- a high skewness which usually disappear with logarithmic transformation,
- a high level of structuration in subpopulations which is generally revealed by multimodality.

This last point has nevertheless some advantages when dealing with clustering of cities.

## REFERENCES

[1] Bagnold, R. and Barndoff-Nielsen, O. (1980). The pattern of natural distribution. *Sedimentology*, 27 (2):199:207.

[2] Batty, M., Carvalho, R., and Hudson-Smith, A. (2008). Scaling and allometry in the building geometries of greater london. *Physics of Condensed Matter*, 63 (3):303:314.

[3] Fieller, N.R.J; Gilbertson, D. and Olbricht, W. (1984). A new method of environmental analysis of particle size distribution data from shoreline sediments. *Nature*, 311 (5987):648:651.