Ingeniería. Revista de la Universidad de Costa Rica
ISSN: 1409-2441
ISSN: 2215-2652
revista.inii@ucr.ac.cr
Universidad de Costa Rica
Costa Rica

Pérez Molina, Eduardo; Vargas Aguilar, Darío

UNCERTAINTY IN LAND VALUE MODELING OF THE SAN JOSÉ METROPOLITAN REGION, COSTA RICA

Ingeniería. Revista de la Universidad de Costa Rica, vol. 34, núm. 1, 2023
Universidad de Costa Rica
Ciudad Universitaria Rodrigo Facio, Costa Rica

DOI: https://doi.org/10.5517/ri.v34i1.56618

Disponible en: https://www.redalyc.org/articulo.oa?id=44176067005

https://revistas.ucr.ac.cr/index.php/ingenieria/index
www.ucr.ac.cr / ISSN: 2215-2652

Ingeniería

Revista de la Universidad de Costa Rica

ENERO/JUNIO 2024 - VOLUMEN 34 (1)

EDITORIAL UCR

UNIVERSIDAD DE COSTA RICA

Ingeniería. Revista de la Universidad de Costa Rica

Vol. 34, No. 1: 43-50, Enero-Junio, 2024. ISSN: 2215-2652. San José, Costa Rica

Ingeniería Civil

Uncertainty in Land Value Modeling of the San José Metropolitan Region, Costa Rica

La incertidumbre en la modelación de valores del suelo de la Gran Área Metropolitana, Costa Rica

Eduardo Pérez Molina [1], *Darío Vargas Aguilar* [2]

[1] *Universidad de Costa Rica, San José, Costa Rica*
correo: eduardo.perezmolina@ucr.ac.cr

[2] *Universidad de Costa Rica, San José, Costa Rica*
correo: dario.vargasaguilar@ucr.ac.cr

Keywords:

Extrapolation, land values, sequential Gaussian simulation, spatial factors, uncertainty.

Abstract

Land value patterns show very distinct spatial associations with accessibility to urban centralities and physical factors in a territory. However, predictions based on models of this structure can be highly uncertain, as the underlying data also may show clustering (thus allowing for better predictions in more densely sampled areas). An assessment of this uncertainty for land value extrapolations in the San José Metropolitan Region of Costa Rica is presented, via conditional Gaussian simulation, and the determinants of this uncertainty were explored, to find spatial strengths and weaknesses in the modeling efforts. The E-Type prediction from the conditional Gaussian simulation was found to marginally improve on ordinary kriging methods and it also provided explicit uncertainty patterns, which are the inverse of the land value prediction. The estimated uncertainty was found to decrease with characteristics that identify suitability for urban land use (and thus higher land values).

Recibido: *12/09/2023*
Aceptado: *30/11/2023*

Palabras Clave:

Extrapolación, factores espaciales, incertidumbre, simulación gaussiana secuencial, valor del suelo.

Resumen

Los patrones de valor del suelo muestran asociaciones espaciales claras con accesibilidad a centralidades urbanas y a factores físicos de un territorio. Sin embargo, las predicciones basadas en esta estructura pueden ser altamente inciertas, dado que los datos mismos también exhiben aglomeración (y, por tanto, permiten mejores predicciones en las zonas más densamente muestreadas). Se presenta una evaluación de esta incertidumbre para extrapolaciones de valor del suelo en la Gran Área Metropolitana de Costa Rica mediante simulaciones gaussianas condicionales y una exploración de los determinantes de esta incertidumbre, como forma de reconocer fortalezas y debilidades de esta predicción. La predicción E-Type simulada resultó marginalmente mejor que extrapolaciones mediante kriging ordinario y produjo una cuantificación espacialmente explícita de la incertidumbre. El patrón de incertidumbre resultó ser un espejo de los valores del suelo. Se encontró que la incertidumbre se reduce con características asociadas a mayor aptitud del suelo para usos urbanos y, por tanto, de mayor precio.

DOI: DOI:10.5517/ri.v34i1.56618

1. INTRODUCTION

The analysis of uncertainty of land value models is a critical issue for policy formulation [1]. However, while the use of Gaussian simulation to understand uncertainty has long been applied to physical land variables (e.g., [2]) and despite kriging having been applied to land rents for at least 20 years [1], no previous cases of conditional Gaussian simulation applied to land value modeling were found .

In general, the analysis of land value in the San José metropolitan region (GAM) has been fragmentary [1]. Recent efforts from extension and research projects at the University of Costa Rica, however, have yielded a data set of real estate listings that provided the first synoptic view of real estate prices in the region [3]. Based on this data, hedonic price models of housing have been produced [3,4] and the first efforts at extrapolation of land values for the entire region (based on kriging and co-kriging) were developed [1]. To isolate land values, [1] consider in their analysis only lots —i.e., properties offered in the land market with no buildings on them and, therefore, with prices only reflecting the attributes of land—; these initial efforts yielded estimates of mean values and of variance, but they were limited to the kriging and co-kriging models.

Given the current state of the question, two objectives are proposed for this paper: first, to extend previous work on land value extrapolation (by [1]) to include conditional Gaussian simulation and, specifically, to include uncertainty estimates for the predicted land value that can be derived with this method; second, to explore whether the uncertainty of these estimates can be explained by spatial structure (indeed, by the same spatial structure related to the point pattern of real estate listings and to the land value pattern itself).

2. METHODOLOGY

2.1 *Land values in the GAM*

Data were compiled by [1] from real estate listings published on the web during 2020-2023. From an original data set of 3670 records with known location and price, extreme values (for the variables price, lot area, and price per square meter) were filtered out, resulting in a final data set of 3196 records. This data set was, in turn, divided by randomly selecting approximately 10 % of records: a calibration data set of 2878 records and a validation data set of 318 records result from this data wrangling process.

The final data sets (of calibration and validation) are shown in Fig. 1. Locations in panel (a) coincide mainly with the urban fabric of the GAM. It is worth noting that the calibration data seem to include a greater proportion of locations in the more central locations (or, conversely, the calibration data are distributed in a way that should better represent the peripheral land values). The land value per square meter was transformed into logarithms (as in [1], [3], [4] and, more generally, following a standard practice in the analysis of land values). The logarithmic transformation of both the calibration and validation data sets (panel (b) of Fig. 1 presents the histogram with a logarithmic scale on the horizontal axis) show a normal distribution, as should be expected, although with some degree of skew towards the left (this may be explained because the filtering process of extreme values is more efficient in excluding excessively large values of price, area, and price per unit of area).

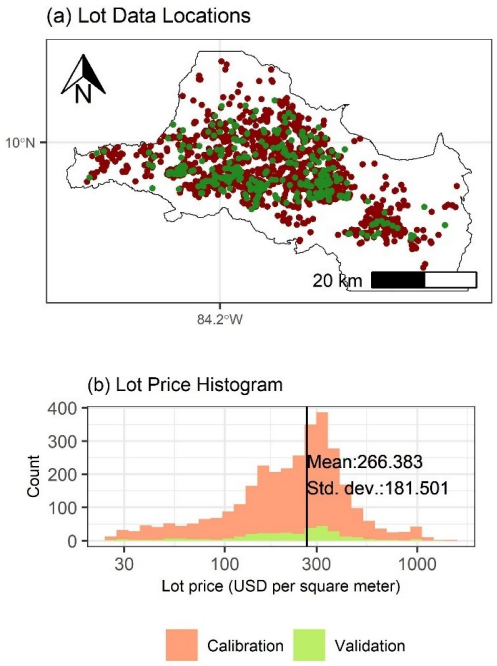


Fig. 1. (a) Lot data locations and (b) histogram of price (log. of USD per square meter) in the GAM (in red, calibration data; in green, validation data).

2.2 *Geostatistical analysis and conditional Gaussian simulation*

As the final objective of the modeling efforts is to produce a spatially explicit prediction of land values per square meter, which is essentially an extrapolation, ordinary kriging was selected to generate a linear weighted estimation for land values at unknown locations from the data set [5], [6], [7].

In kriging, the spatial dependence structure is modeled through a semivariogram, a function that relates the mean semivariance (the squared differences in the Z value for pairs of locations xi, for locations with known Z values) for all N pairs of locations within a range of distances h [1],[6],[7]:

The empirical semivariogram is fitted by a function with a specified form; for the GAM, [1] proposed a spheric adjustment; the gstat package can determine the optimal parameters for this function, based on the data [8].

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x_i + h) - Z(x_i)]^2 \quad (1)$$

Under the kriging method, the predicted $\hat{Z}$ values for x_i locations with unknown land values result from a weighted average of locations (with weights w_i) with known land value, $Z(x_i)$:

$$D = sup_x|F_m(x) - G_n(x)| \quad (2)$$

Ordinary kriging chooses the optimal weights by minimizing kriging variance, which can be determined from the semivariogram model [5],[6].

The uncertainty estimate for the extrapolation was determined using sequential Gaussian simulation, which is a technique to systematically simulate realizations of a random field. Given a semivariogram from the data and a random path through all locations with no known values (such that each location is only visited once), the sequential Gaussian simulation algorithm proceeds as follows: (1) it searches for all sampled data and for all previously simulated locations, (2) it applies kriging to neighboring points and determines from it the linear estimate and its variance, (3) sample the value from a normal distribution with mean and variance from the kriging of the previous step, (4) assign the sampled value to the location and proceed along the random path to the next location [5]. The simulation is termed conditional because it is conditioned on the data, via the kriging.

One hundred instances of land value patterns were simulated using conditional Gaussian simulation; at each instance, only the closest 320 points to the location being extrapolated (approximately 10 % of the calibration data) were considered in the kriging. Data on each simulated instance were back-transformed into their original units. Based on these simulations, the following metrics were reported: (1) the E-Type prediction, which is the per location (i.e., ensemble) mean of the land value [5], (2) the per location standard deviation and coefficient of variation (the coefficient of variation is the ratio of standard deviation to mean), and (3) the per location 95th and 5th percentiles, as plausible bounds within which the actual land value should be found. Calculations were performed using the gstat package [8] of statistical software R [9].

2.3 *The determinants of uncertainty: an exploration of social and physical factors*

The pattern of the simulated standard deviations was analyzed to explore its association with other possible spatial factors, in line with the objectives proposed. The variance follows a χ^2 distribution. Therefore, to find the statistical significance of the variation in the standard deviation associated to any given factor, the following approach was employed: for all locations in the prediction space, (1) histograms were computed for each spatial factor and the locations were classified into three separate groups based on limits defined by changes in the histograms of the spatial factor; (2) a Kolmogorov-Smirnov non-parametric test was conducted to determine whether the statistical distribution of the standard deviation of any group was different from each of the other groups (for each factor separately); (3) smooth kernel densities were estimated for each group (using the *geom_density()* function of the package ggplot [10] from R [9]); these densities were compared. The relative positioning of the different kernel densities (determined by the group) was interpreted to understand how the factor affected the standard deviation. The general expectation was that factors associated with greater suitability for urban land use such as flatter terrain and greater accessibility would present less uncertainty, as they should also be correlated with greater density of locations with known land values [11].

Following [12], the Kolmogorov-Smirnov test is the most common instrument to explore the hypothesis of whether two samples are taken from the same statistical distribution. Taking two samples, $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$ from two distribution functions, F and G, one may form the empirical distribution functions F_m: =| {xi: xi ≤x} |/m and G_n: =| {yi: yi ≤y} |/n. The test statistic for the null hypothesis that $F = G$ is given by:

$$\hat{Z} = \sum_{i=1}^{n} w_i Z(x_i) \quad (3)$$

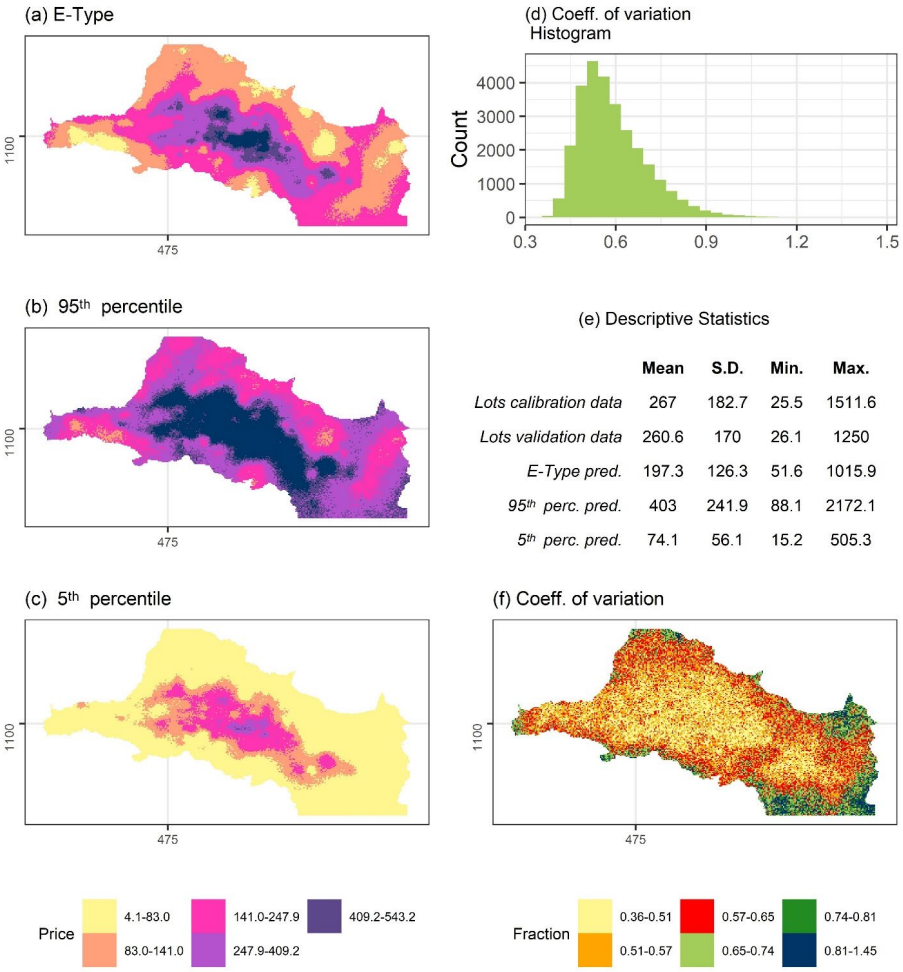
which should be contrasted using probabilities from the cumulative Kolmogorov distribution [12].

3. RESULTS

As was described, 100 instances of the logarithm of land values in the GAM were simulated (their back-transformed mean, 95th and 5th percentiles are reported in Fig. 2).

Two important findings can be seen in Fig. 2, perhaps the most important is reported in panels.

(d), (e), and (f): these summarize the uncertainty of the E-Type estimates. The pattern of the coefficient of variation is, approximately, the reverse of the land value predictions (most clearly seen when compared with the E-Type prediction of panel (a)). When contrasted with the density of points (Fig. 1, panel (a)), there is also a clear association. However, the effect of the algorithm can be seen in that the yellow area of low coefficient of variation extends beyond the more urbanized area predicted by large land values (the darker blue and purple intervals of Fig. 2, panel (a), which are also the areas with most sale listings in Fig. 1, panel (a)). The coefficient of variation mostly predicts standard deviations varying between 36 % and 65 % of the mean, suggesting adequately precise measurements (relatively low dispersion in the simulated instances); their distribution tends to be right skewed within this range (shown in the histogram of Fig. 2, panel (d)).



	Mean	S.D.	Min.	Max.
Lots calibration data	267	182.7	25.5	1511.6
Lots validation data	260.6	170	26.1	1250
E-Type pred.	197.3	126.3	51.6	1015.9
95th perc. pred.	403	241.9	88.1	2172.1
5th perc. pred.	74.1	56.1	15.2	505.3

Fig. 2. Conditional Gaussian Simulation of land value in the GAM (USD per square meter). (a) E-Type prediction (cell-wise mean of simulated instances), (b) 95th percentile of simulated instances, (c) 5th percentile of simulated instances, (d) location-wise coefficient of variation histogram, (e) descriptive statistics of simulated predictions and data, (f) coefficient of variation map.

The second relevant finding of Fig. 2 corresponds to the interpretation of panels (a), (b), and (c): in effect, the E-Type prediction (of panel (a)) is the best estimate of land value; panels (b) and (c) represent higher and lower bound values for this prediction: the land value for 90 % of simulated instances was estimated to lie within the range for each location. An examination of this more detailed set of maps suggests uncertainties may be larger than what the overall measures (of validation, discussed and reported in TABLE I, and of the coefficient of variation) had suggested. By comparing location-wise, for most locations, the upper (95th percentile) or lower bound (5th percentile) shift one category. Given the values involved, this represents around double the back-transformed mean value.

It is also important to understand the quality of the predictions. TABLE I summarizes the validation exercise results. Following the methodological approach, 10 % of the lots data were reserved for validation. For these locations, predictions of land value were generated using (a) the E-Type prediction of the conditional Gaussian simulation (the location-wise mean of all instances) and (b) an ordinary kriging extrapolation, as benchmark. The error was calculated by subtracting the land value per square meter (of the data set) from the back-transformed predicted value. Examining their absolute values, in general, the error terms showed both models underestimated the actual land value.

As can be seen in TABLE I, the E-Type land value prediction is slightly worse than the ordinary kriging: all estimates of error of the E-Type prediction are somewhat larger than the corresponding value for ordinary kriging, as is the range is also smaller. The differences are very small, in general (at least an order of magnitude smaller than the error estimate). It is also worth pointing out that both models have produced very accurate

predictions: all mean and median error and RMSE estimates are all less than half of the variable mean (for the land value per square meter of the validation data set, which is reported in Fig. 2).

The predicted pattern of land values per square meter is shown in Fig. 2. The pattern coincides with theoretical expectations and indeed with previous, kriging-based analysis of land values in the GAM from [1]: land values are larger (shown in dark blue color) for the centers of San José and Heredia, and the centers of Alajuela and Cartago are also relatively larger than their surroundings. Furthermore, lower values are concentrated on the periphery of the region (rural areas) and the northern zones of Alajuela and Heredia tend to exhibit larger values than those of Cartago and San José.

The second objective of this paper, following the estimation of uncertainties, is to explore if these uncertainties respond to regularities in space. To do so, five factors that determine suitability for urban development (and, in consequence, are related to land price formation in urban markets) were considered: slope and elevation, and (Euclidean) distance to the CBD, to the nearest municipal center and to the nearest main road. For each factor, three groups of locations were created (except for elevation, for which only two groups were defined) based on the factor value; group intervals were generally defined based on the variable histograms, although for slope, the group limits are related to statutory building requirements.

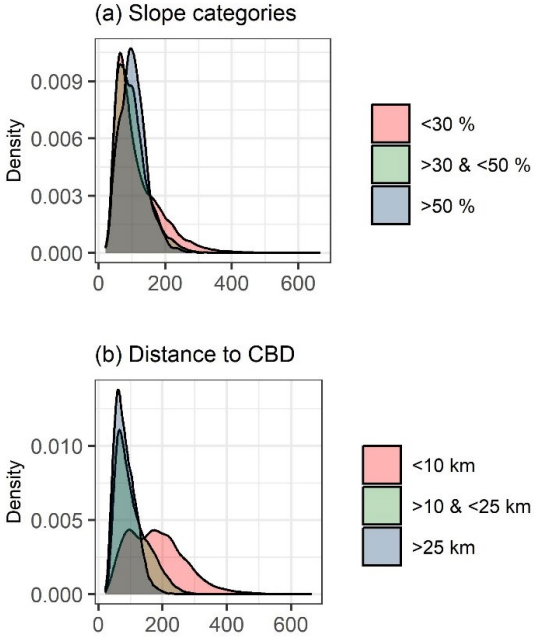
TABLE I
VALIDATION OF PREDICTION MODELS OF LAND PRICE (USD per square meter)

Error measure	Prediction model	
	Conditional Gaussian Simulation (E-Type)	Ordinary Kriging
Root Mean Square Error	149.4	127.7
Mean Absolute Error	110.7	84.3
Median Absolute Error	82.5	56.8
Range of Error	-569.2 – 677.0	-437.3 – 765.1

TABLE II
KOLMOGOROV-SMIRNOV STATISTICS FOR DISTRIBUTION OF STANDARD DEVIATION FOR DEVELOPED PREDICTIONS GROUPED BY DETERMINANTS

Comparison	D Statistic	Prob.
Slope		
G1: <30 % *vs. G2*:>30 % & <50 %	0.110	<0.01
G1: <30 % *vs. G3*:>50 %	0.137	<0.01
G2:>30 % & <50 % vs. *G3*:>50 %	0.117	<0.01
Elevation		
G1: <1500 masl *G2*: >1500 masl	0.193	<0.01
Distance to CBD		
G1: <10 km *vs. G2:* >10 km & < 25 km	0.412	<0.01
G1: <10 km *vs. G3*: > 25 km	0.583	<0.01
G2: >10 km & < 25 km vs. *G3:* > 25 km	0.178	<0.01
Distance to nearest municipal center		
G1: <2.5 km *vs. G2:* >2.5 km & < 7.5 km	0.364	<0.01
G1: <2.5 km *vs. G3:* > 7.5 km	0.362	<0.01
G2: >2.5 km & < 7.5 km vs. *G3:* > 7.5 km	0.120	<0.01
Distance to nearest main road		
G1: <1 km vs. *G2:* >1 km & < 7.5 km	0.268	<0.01
G1: <1 km vs. *G3:* > 7.5 km	0.409	<0.01
G2: >1 km & < 7.5 km vs. *G3:* > 7.5 km	0.160	<0.01

The Kolmogorov-Smirnov test was used to explore whether the statistical distribution of standard deviation for each group was different from other groups for the same factor. These results are shown in TABLE II and, as should have been expected, all test statistics confirmed the distribution of data of one group is distinct from other groups. On the one hand, there are sufficient simulated locations (over 28000) for even small differences to be significant. On the other, the larger probability of urbanization associated with flatter zones with greater accessibility to urban centralities is also associated with both the point pattern of real estate sales listings [11] –i.e., the sampling density, a key determinant of uncertainty—and the land value itself.



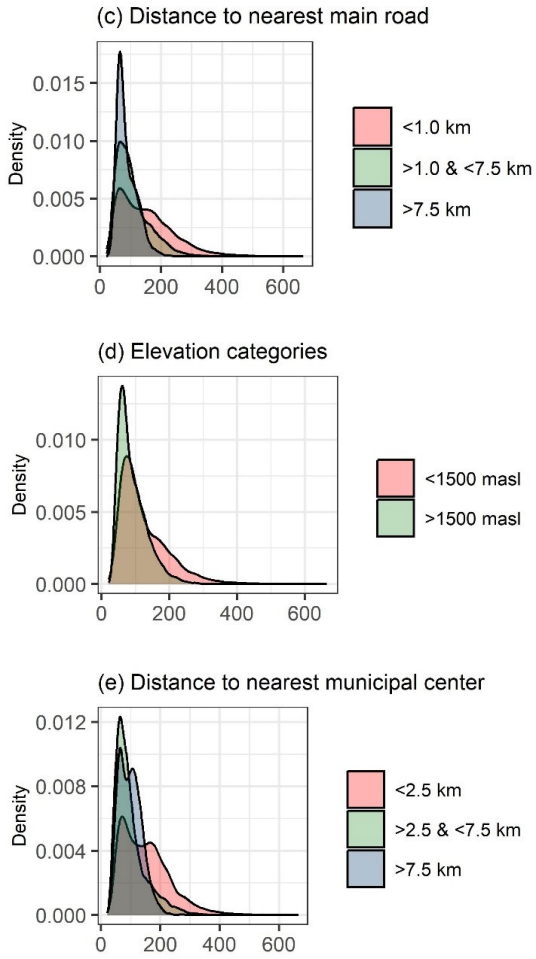


Fig. 3. Empirical cumulative distribution functions for standard deviation of simulated predictions (square of the log. of USD per square meter). Locations grouped by (a) slope, (b) Euclidean distance to CBD, (c) Euclidean distance to main roads, (d) elevation, and (e) Euclidean distance to nearest municipal center.

How each factor affects uncertainty (measured by the standard deviation of land value of the simulated instances) suggests urban areas have more diverse land values than zones less suitable for urban uses. Fig. 3 shows kernel smoothed empirical distribution densities for the location-wise standard deviation of simulated instances, grouped by the categories that were used in constructing TABLE II. The steeper locations (slopes greater than 50 %) have distinctly larger uncertainty (a sharper peak at higher value of the distribution) than other groups. This same pattern is repeated for all variables: greater accessibilities to urban centralities (the CBD, the nearest municipal center) or the regional transportation network (main roads), represented by the pink density function estimate, have all lower peaks at the lower end of the standard deviation values, suggesting more dispersed values. In general, the intermediate group of factor values (shown in light green, Fig. 3) presents intermediate levels of uncertainty and the group of larger factor values (light blue, Fig. 3), lower levels of uncertainty (the density functions for intermediate groups are less right skewed than those for the larger groups of factor values).

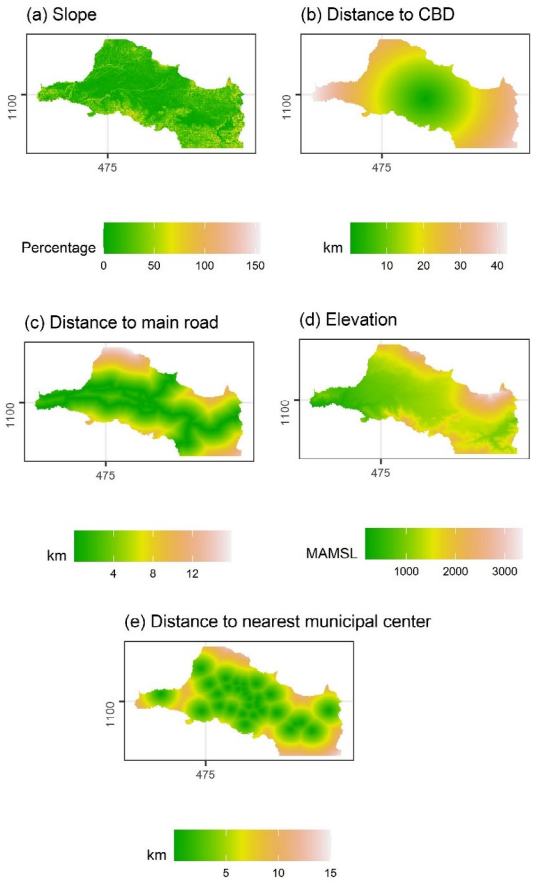


Fig. 4. Spatial patterns of determinants of uncertainty. (a) Slope, (b) Euclidean distance to CBD, (c) Euclidean distance to main roads, (d) elevation, and (e) Euclidean distance to nearest municipal center.

4. SYNTHESIS AND DISCUSSION

The analysis of land value patterns extended previous results and it has provided further insights, which have contributed to identify both needs for further study and opportunities for applications to public policy.

The E-Type prediction from the conditional Gaussian simulation was found to marginally improve on ordinary kriging methods. The conditional Gaussian simulation produced, for validation data, slightly better error measures (RMSE, mean, median, and range of error) than ordinary kriging (in the analysis of variations in kriging methods conducted by [1], the different methods tested also resulted in very similar error levels for validation). This result is indeed not surprising, as the simulations are conditional on the variogram, and should more iterations had been simulated, the difference would have likely been smaller. On the other hand, in so far as improvements were generated by the simulations, they were likely related to the improvement of

over-smoothing limitations in the kriging predictions [5]; but even this feature could have likely been incorporated into the kriging by a careful consideration on the number of neighboring points determining a prediction. Previous exercises of kriging models did find limitations due to this over-smoothing problem that seem to have been improved on by the sequential Gaussian simulation method (in particular, by better modeling the local changes at the peri-urban interface of the region); further work on this issue seems promising.

A distinct advantage of conditional Gaussian simulation is the spatially explicit measures of uncertainty that can be used to explore the limitations of the prediction and to more easily estimate exceedance probabilities [13]; this feature is especially useful for land value maps in applied scenarios (for example, when the map predicts land values claimed to be too large by a land owner, this claim can be easily tested). Further work is required on this issue (previous comparisons of models estimated from this data and other data sources suggest systematic underestimation of land values, particularly for taxation purposes [3]; while outdated assessments are the simplest explanation, it is also possible that data sources for the models reported in this paper may be also partially skewing the results).

It is further worth noting that the literature has detected over-smoothing problems associated with deterministic methods such as ordinary kriging that can be overcome with simulation. The current focus of this study was not the comparison of conditional Gaussian simulation with other extrapolation predictions; however, this is regarded as a potential area for further investigation.

The estimated uncertainty patterns are inversely related to the predicted land value. A very clear and negative spatial association was identified between the E-Type prediction of land values per square meter and its standard deviation: in the urban central area of the GAM, the highest land values (which coincides both with previous analysis [1] and with theoretical expectations from urban economics) and lowest uncertainties were observed. This finding coincides with previous analysis of the point pattern of real estate listings and its relation to the determinants of suitability for urban land uses [11].

Indeed, the estimated uncertainty was found to decrease with characteristics that identify suitability for urban land use (and thus higher land values). The flatter areas of the GAM, which are also closer to urban centralities (the CBD, main municipal centers), showed much less uncertainty (smaller location-wise standard deviation) than zones further away and at higher elevations and steeper slopes. Therefore, the data set and modeling efforts appear to demonstrate efficiency when predicting urban land values but also present clear limitations if applied to rural land uses of the urban periphery.

Despite its importance, hardly any previous case study reports the use of simulation to understand uncertainty introduced by interpolation into land or property value predictions (unlike physical properties of soils, which are derived from similar point data and for which such analysis seems common). Uncertainty has been reported as variance of kriging estimates [1] or verification through out-of-sample prediction [14], in relation to the mean estimate from this indicator. While theoretical recognition of the possibility to estimate errors and uncertainty in the context of land valuation has been acknowledged [15], actual practice has centered on the accuracy of the mean prediction rather than on explaining its variance. Uncertainty is important for valuation, especially when practical applications are performed (such as tax assessments and potential challenges to these).

In conclusion, the analysis of uncertainties may be critical for improving urban and regional studies (e.g., the impact of new infrastructure or of land use regulations) and land value assessments for tax policy. In this regard, the methods presented have increased robustness (relative to very local estimates) because predictions relatively far away from locations with known values may still benefit from their price information via the spatial dependence encoded in the semivariogram. More importantly, the estimates of uncertainty permit the assessment of the prediction for properties that have not been recently sold in the market (and thus include an inherent check of the prediction which is absent in isolated tax assessment exercises).

ROLES

Eduardo Pérez Molina: Conceptualization, Methodology, Software, Formal analysis, Writing - Original Draft

Darío Vargas Aguilar: Data Curation, Writing - Review & Editing

REFERENCES

[1] E. Pérez-Molina and M. Román-Forastelli, “Análisis geo-estadístico de los patrones de valores del suelo en la Gran Área Metropolitana de Costa Rica,” manuscript submitted for publication, 2023.

[2] D.H. Easly, L.E. Borgman, and P.N. Shive, “Geostatistical simulation for geophysical applications. Part I: simulation,” *Geophysics*, vol. 55, no. 11, pp. 1435-1440, 1990, doi: 10.1190/1.1442790.

[3] M. Román, A. Quirós, A., E. Pérez *et al*., “Dinámica inmobiliario y formación de precios: primer mapa de valores de mercado del suelo en la GAM,” Proyecto ED-3466 Asesoría técnica y económica a la política pública. Escuela de Economía, Universidad de Costa Rica, San José, Costa Rica, 2023.

[4] E. Pérez-Molina, “Exploring a multilevel approach with spatial effects to model housing price in San José, Costa Rica,” *Env & Plan*, vol. 49, no. 3, pp. 987-1004, 2022, doi: 10.1177/23998083211041122.

[5] T. Bai and P. Tahmasebi, “Sequential Gaussian simulation for geosystems modeling: A machine learning approach,” *Geosci Front*, vol. 13, pp. 101258, 2022, doi: 10.1016/j.gsf.2021.101258.

[6] N.A.C. Cressie, *Statistics for Spatial Data.* Wiley, New York, 1993.

[7] Goovaerts, *Geostatistics for Natural Resources Evaluation.* Oxford, United Kingdom: Oxford University Press, 1997.

[8] E.J. Pebesma, "Multivariable geostatistics in S: the gstat package," *Comput & Geosci*, vol. 30, pp. 683-691, 2004, doi: 10.1016/j.cageo.2004.03.012.

[9] R Core Team, "R: A Language and Environment for Statistical Computing," R Foundation for Statistical Computing. Vienna, Austria, 2023.

[10] H. Wickham, *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag, New York, 2016.

[11] E. Pérez-Molina, "Understanding the spatial statistical properties of a real estate listings point pattern in San José, Costa Rica," manuscript submitted for publication, 2023.

[12] T. Viehmann, "Numerically more stable computation of the p-values for the two-sample Kolmogorov-Smirnov test," *arXiv preprint*, arXiv:2102.08037, 2021.

[13] S. Metahni, L. Coudert, E. Gloaguen et al., "Comparison of different interpolation methods and sequential Gaussian simulation to estimate volumes of soil contaminated by As, Cr, Cu, PCP and dioxins/furans," *Environ Pollut*, vol. 252, pp. 409-419, 2019, doi: 10.1016/j.envpol.2019.05.122

[14] K. de Koening, T. Filatova, and O. Bin, "Improved Methods for Predicting Property Prices in Hazard Prone Dynamic Markets," *Environ Resource Econ*, vol. 69, pp. 247–263, 2018, doi: 10.1007/s10640-016-0076-5

[15] R. Cellmer, "The Possibilities and Limiations of Geostatistical Methods in Real Estate Market Analysis," *Real Estate Manag*, vol. 22, pp. 54-62, doi: 10.2478/remav-2014-0027

Uncertainty of land values in the GAM

Eduardo Pérez Molina* & Darío Vargas Aguilar

2023-09-05

```
knitr::opts_chunk$set(echo = TRUE)
suppressMessages(library(sf))
suppressMessages(library(gstat))
suppressMessages(library(dplyr))
suppressMessages(library(tidyverse))
suppressMessages(library(sp))
suppressMessages(library(ggplot2))
suppressMessages(library(ggspatial))
suppressMessages(library(stars))
suppressMessages(library(grid))
suppressMessages(library(gtable))
suppressMessages(library(gridExtra))
```

This document documents an exploration of uncertainty's determinants when modeling land values. We begin by loading the final data from our extrapolation paper (which include land value records to date for 2020, 2021, 2022, and 2023, already having filtered out the extreme values).

```
load("G:/20230815/16_ValorSueloIncertidumbre/01_Datos/MarkDown_all_20230722_papergraphs.RData")
#Data set available upon request from corresponding author
rm(list=setdiff(ls(), c("GAM","lcu_pol","lots","lots1","lots2","st_grid")))
```

We will now proceed as follows: (1) we generate 100 simulations of the land value pattern using sequential Gaussian simulation and we use these to estimate the uncertainty in the data and (2) we then test whether the uncertainty estimates vary systematically for a group physical factores (slope, elevation, distance to CBD).

Sequential Gaussian Simulation

We begin by fitting the semivariogram to the lots data (we use the full dataset *lots1*). Recall we have a good approximation to starting values from the kriging paper we had previously written:

```
v.lots <- variogram(log(unitprice) ~ 1, lots1)
```

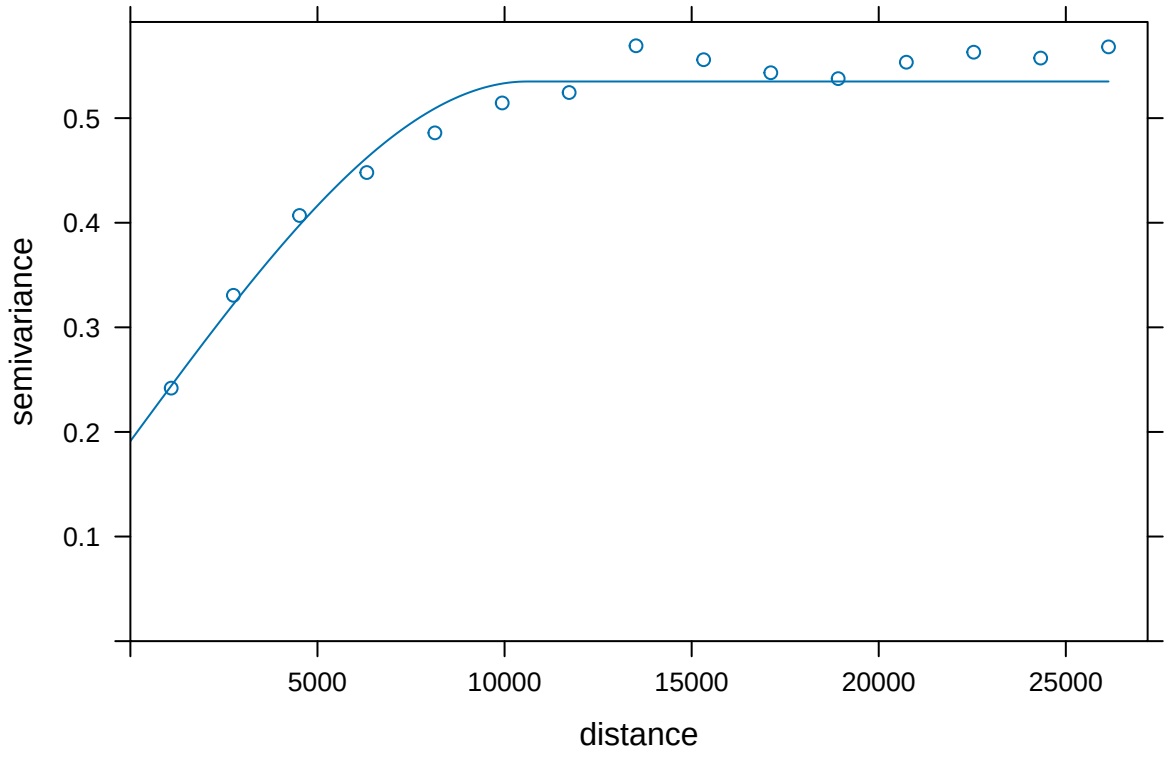
```
## Please note that rgdal will be retired during October 2023,
## plan transition to sf/stars/terra functions using GDAL and PROJ
## at your earliest convenience.
## See https://r-spatial.org/r/2023/05/15/evolution4.html and https://github.com/r-spatial/evolution
## rgdal: version: 1.6-7, (SVN revision 1203)
## Geospatial Data Abstraction Library extensions to R successfully loaded
## Loaded GDAL runtime: GDAL 3.6.2, released 2023/01/02
## Path to GDAL shared files: C:/Users/Eduardo/AppData/Local/R/win-library/4.3/rgdal/gdal
##  GDAL does not use iconv for recoding strings.
## GDAL binary built with GEOS: TRUE
## Loaded PROJ runtime: Rel. 9.2.0, March 1st, 2023, [PJ_VERSION: 920]
```

```
## Path to PROJ shared files: C:/Users/Eduardo/AppData/Local/R/win-library/4.3/rgdal/proj
## PROJ CDN enabled: FALSE
## Linking to sp version:2.0-0
## To mute warnings of possible GDAL/OSR exportToProj4() degradation,
## use options("rgdal_show_exportToProj4_warnings"="none") before loading sp or rgdal.
```

```
m.lots <- vgm(0.4, "Sph", 10000, 0.2)#Starting model
(m.lots <- fit.variogram(v.lots, m.lots))
```

```
##   model     psill    range
## 1   Nug 0.1912351     0.00
## 2   Sph 0.3437995 10605.57
```

```
plot(v.lots, pl=F, model=m.lots)
```



Before coding the simulation, we create descriptives for the input data:

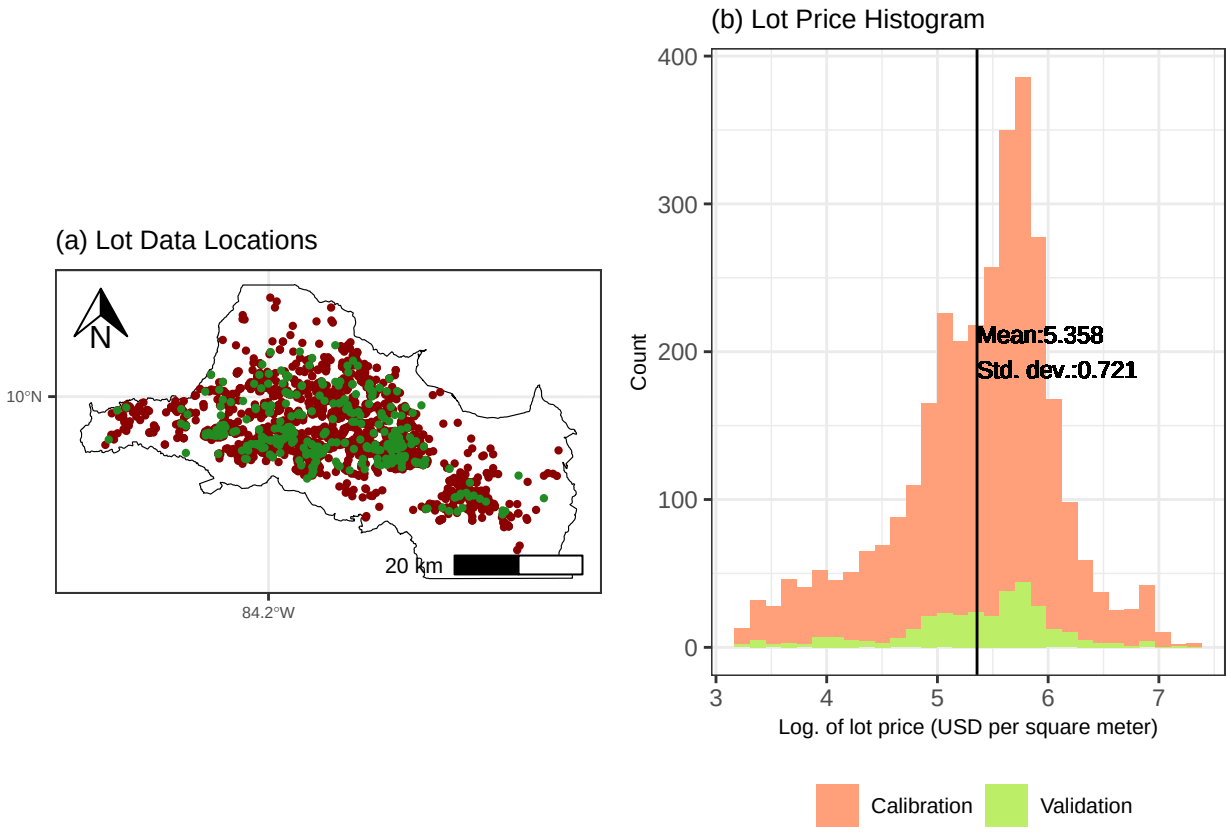
```
temp <- rbind(cbind(lots1@data[,1:19],Submuestra=rep(1,length(lots1$X))),
              cbind(lots2@data[,1:19],Submuestra=rep(2,length(lots2$X))))
#colnames(temp@data)[20] <- "Submuestra"
temp$Submuestra <- as.factor(temp$Submuestra)

plot1 <- ggplot() +
  layer_spatial(GAM, col="black",fill=NA)+
  layer_spatial(lots1, col="red4",size=0.8)+
  layer_spatial(lots2, col="forestgreen",size=0.8)+
  labs(title="(a) Lot Data Locations") +
  geom_sf() +
```

```
  scale_y_continuous(breaks = c(10)) +
  scale_x_continuous(breaks = c(-84.2)) +
  theme_bw() + theme(legend.position="bottom",
                     legend.text=element_text(size=8),
                     legend.title=element_text(size=8),
                     text = element_text(size = 8)) +
  annotation_north_arrow(location="tl", width = unit(.7, "cm"),
                         height = unit(.7, "cm")) +
  annotation_scale(location="br")
plot2 <- ggplot(data=as.data.frame(temp),aes(x=log(unitprice))) +
  geom_histogram(aes(fill=Submuestra)) +
  scale_fill_manual(values = c("lightsalmon1","darkolivegreen2"),labels=c("Calibration","Validation")) +
  geom_vline(xintercept = mean(log(temp$unitprice))) +
  geom_text(aes(x=mean(log(temp$unitprice)), y=200,
                label=paste("Mean:",round(mean(log(temp$unitprice)),3),"\nStd. dev.:",round(sd(log(temp$
            hjust=0, size=3.2) +
  ggtitle("(b) Lot Price Histogram") +
  xlab("Log. of lot price (USD per square meter)") + ylab("Count") + labs(fill="") +
  theme_bw() + theme(legend.position="bottom",
                     legend.text=element_text(size=8),
                     legend.title=element_text(size=8),
                     axis.title=element_text(size=8),
                     plot.title = element_text(size=9.5))
```

And the figure for the text:

```
#jpeg(filename="F:/20230815/16_ValorSueloIncertidumbre/03_Reporte/Figuras/Fig01_location.jpg",
#     width=6.5,height=2.2,units="in",res=600)
grid.arrange(plot1,plot2,ncol=2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#dev.off()
rm(temp)
```

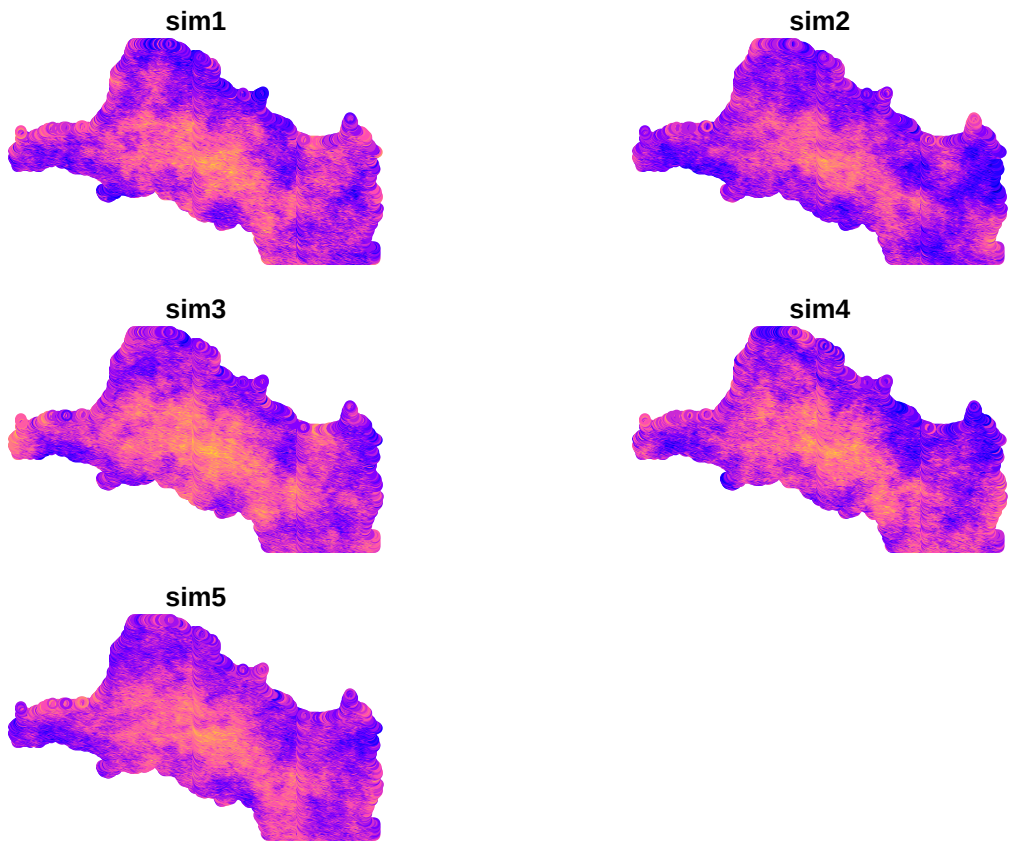
We now test for 5 simulations (following Pebesma's example):

```
set.seed(131)
sim <- krige(formula = log(unitprice)~1, lots1, st_grid, model = m.lots,
  nmax = 290, nsim = 5)
```

```
## drawing 5 GLS realisations of beta...
## [using conditional Gaussian simulation]
```

We now look at the five simulations:

```
plot(sim)
```

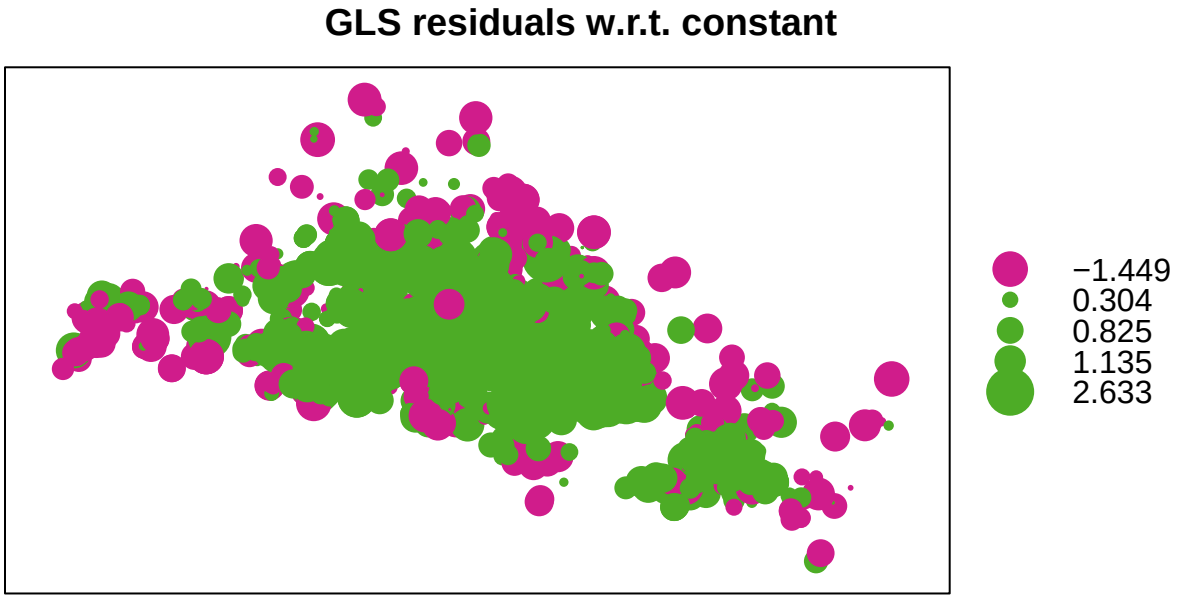


Further explorations: the residuals.

```
# calculate generalised least squares residuals w.r.t. constant trend:
g <- gstat(NULL, "log.unitprice", log(unitprice)~1, lots1, model = m.lots)
blue0 <- predict(g, newdata = lots1, BLUE = TRUE)
```

```
## [generalized least squares trend estimation]
```

```
#> [generalized least squares trend estimation]
blue0$blue.res <- log(lots1$unitprice) - blue0$log.unitprice.pred
bubble(blue0, zcol = "blue.res", main = "GLS residuals w.r.t. constant")
```



And we are now ready to estimate 100 instances:

```
set.seed(1012)
sim.full <- krige(formula = log(unitprice)~1, lots1, st_grid, model = m.lots,
                  nmax = 290, nsim = 100)
```

```
## drawing 100 GLS realisations of beta...
## [using conditional Gaussian simulation]
```

Compute uncertainty measures

We now estimate mean, standard deviation, and variation coefficient of the simulations. To do so, we convert the simulation into a data frame (with the coordinates as variables), we estimate mean, standard deviation, and coefficient of variation, and we transform back into an sf object.

```
separated_coord <- sim.full %>%
    mutate(long = unlist(map(sim.full$geometry,1)),
           lat = unlist(map(sim.full$geometry,2)))
sim.full <- as.data.frame(sim.full)
```

The estimates for mean, standard deviation, coefficient of variation:

```
for(i in 1:length(sim.full[,1])){
  sim.full$mean[i] <- mean(as.numeric(sim.full[i,1:100]),na.rm=TRUE)
  sim.full$sd[i] <- sd(as.numeric(sim.full[i,1:100]),na.rm=TRUE)
  sim.full$cv[i] <- sim.full$sd[i]/sim.full$mean[i]
  #print(sim.full$mean[i])
}
```

```
rm(i)
```

And now we re-cast the result into an sf object:

```
sim.full <- st_as_sf(sim.full, crs = 5367)
```

Describe uncertainty patterns and validate model

Validation with *lots2* dataset

For every data point in *lots2*, we assign to it the value of the nearest point in the simulation grid (and therefore all values from the simulation):

```
lots2 <- st_join(st_as_sf(lots2),st_as_sf(sim.full),join=st_nearest_feature)
lots2 <- as.data.frame(lots2)
for(i in 1:length(lots2[,1])){
  lots2$mean[i] <- mean(as.numeric(lots2[i,25:124]),na.rm=TRUE)
  lots2$sd[i] <- sd(as.numeric(lots2[i,25:124]),na.rm=TRUE)
  lots2$cv[i] <- lots2$sd[i]/lots2$mean[i]
  #print(sim.full$mean[i])
}
rm(i)
```

We then use the E-type estimate as the predicted value and then calculate estimates of error to validate the simulation model:

```
error.val <- log(lots2$unitprice)-lots2$mean
mean(abs(error.val))#mean error
```

```
## [1] 0.3792764
```

```
median(abs(error.val))#median error
```

```
## [1] 0.2601104
```

```
sqrt(sum(error.val^2)/length(error.val))#RMSE (precision)
```

```
## [1] 0.5275186
```

```
range(error.val)#Range of estimated error
```

```
## [1] -2.779020  1.841764
```

We note the errors (of the logged unit price) are small: the mean of the logged unit price in *lots2* is 5.349; by comparison, mean and median errors are less than a 10th of it and the RMSE is of similar order of magnitude (despite being expressed as the square of errors). So the model is quite good. (In the paper: compare to the O.K. from the extrapolation piece.)

We further estimate the error measures for an ordinary kriging model, as benchmark:

```
ok.err <- krige(log(unitprice) ~ 1,
          locations=lots1,
          newdata=st_as_sf(lots2, crs = 5367),
          model=m.lots)
```

```
## [using ordinary kriging]
```

```
error.val.ok <- log(lots2$unitprice)-ok.err$var1.pred
mean(abs(error.val.ok))#mean error
```

```
## [1] 0.3722722
```

```
median(abs(error.val.ok))#median error
```

```
## [1] 0.2474721
```

```
sqrt(sum(error.val.ok^2)/length(error.val.ok))#RMSE (precision)
```

```
## [1] 0.5221011
```

```
range(error.val.ok)#Range of estimated error
```

```
## [1] -2.708782  1.805745
```

Uncertainty simulation results

We now examine: (1) the maps of E-type, C.V., plus the 5th and 95th percentiles, (2) the histogram of C.V.

We begin by estimating the 5th and 95th percentiles:

```
sim.full <- as.data.frame(sim.full)
for(i in 1:length(sim.full[,1])){
  sim.full$q05[i] <- quantile(as.numeric(sim.full[i,1:100]), probs=0.05)
  sim.full$q95[i] <- quantile(as.numeric(sim.full[i,1:100]), probs=0.95)
}
rm(i)
```

We now determine the categories for the map colors and create categorical variables for mapping:

```
#For predictions:
quantile(as.matrix(sim.full[,1:100]), probs=c(0,0.25,0.5,0.75,0.9,0.95,1))
```

```
##       0%      25%      50%      75%      90%      95%     100%
## 1.212987 4.218608 4.748454 5.312895 5.813971 6.097342 8.299906
```

```
#For C.V.
quantile(sim.full$cv, probs=c(0,0.25,0.5,0.75,0.9,0.95,1))
```

```
##         0%        25%        50%        75%        90%        95%       100%
## 0.05791793 0.09663170 0.11754878 0.13754983 0.15184679 0.15883800 0.19216319
```

```
#Categories for E-type
sim.full$CatMean <- 1
sim.full$CatMean[sim.full$mean>4.220160] <- 2
sim.full$CatMean[sim.full$mean>4.753129] <- 3
sim.full$CatMean[sim.full$mean>5.315921] <- 4
sim.full$CatMean[sim.full$mean>5.813729] <- 5
sim.full$CatMean[sim.full$mean>6.095221] <- 6
#Categories for 5th
sim.full$Catq05 <- 1
sim.full$Catq05[sim.full$q05>4.220160] <- 2
sim.full$Catq05[sim.full$q05>4.753129] <- 3
sim.full$Catq05[sim.full$q05>5.315921] <- 4
sim.full$Catq05[sim.full$q05>5.813729] <- 5
sim.full$Catq05[sim.full$q05>6.095221] <- 6
#Categories for 95th
sim.full$Catq95 <- 1
sim.full$Catq95[sim.full$q95>4.220160] <- 2
sim.full$Catq95[sim.full$q95>4.753129] <- 3
sim.full$Catq95[sim.full$q95>5.315921] <- 4
sim.full$Catq95[sim.full$q95>5.813729] <- 5
```

```
sim.full$Catq95[sim.full$q95>6.095221] <- 6
#Categories for CV
sim.full$Catcv <- 1
sim.full$Catcv[sim.full$cv>0.09628661] <- 2
sim.full$Catcv[sim.full$cv>0.11665832] <- 3
sim.full$Catcv[sim.full$cv>0.13671871] <- 4
sim.full$Catcv[sim.full$cv>0.15141502] <- 5
sim.full$Catcv[sim.full$cv>0.15875727] <- 6
#Check
table(sim.full$CatMean)
```

```
##
##     1     2     3     4     5     6
##  4120 12350  6171  4308  1011   397
```

```
table(sim.full$Catq05)
```

```
##
##     1     2     3     4     5
## 20185  4575  3138   446    13
```

```
table(sim.full$Catq95)
```

```
##
##     2     3     4     5     6
##   556  7738 10293  3738  6032
```

```
table(sim.full$Catcv)
```

```
##
##    1    2    3    4    5    6
## 6959 6883 7122 4452 1505 1436
```

```
sim.full <- st_as_sf(sim.full, crs = 5367)
```

And the summary statistics table:

```
tab.var <- rbind(c(round(mean(log(lots1$unitprice)),3),round(sd(log(lots1$unitprice)),3),
                   round(min(log(lots1$unitprice)),3),round(max(log(lots1$unitprice)),3)),
                 c(round(mean(log(lots2$unitprice)),3),round(sd(log(lots2$unitprice)),3),
                   round(min(log(lots2$unitprice)),3),round(max(log(lots2$unitprice)),3)),
                 c(round(mean(sim.full$mean),3),round(sd(sim.full$mean),3),
                   round(min(sim.full$mean),3),round(max(sim.full$mean),3)),
                 c(round(mean(sim.full$q95),3),round(sd(sim.full$q95),3),
                   round(min(sim.full$q95),3),round(max(sim.full$q95),3)),
                 c(round(mean(sim.full$q05),3),round(sd(sim.full$q05),3),round(min(sim.full$q05),3),rou
tab.var <- as.data.frame(tab.var)
colnames(tab.var) <- c("Media","Desv. est.","Mín.","Máx.")
rownames(tab.var) <- c("Lots calibration data","Lots validation data","E-Type pred.",
                       "95th perc. pred.","5th perc. pred.")
```

To show it in the figure, we must convert it to a grob object:

```
panel6 <- tableGrob(tab.var, theme=ttheme_minimal(core = list(fg_params=list(cex = 0.7)),colhead = list
title <- textGrob("(e) Descriptive Statistics\n",gp=gpar(fontsize=9))
panel6 <- gtable_add_rows(panel6,heights = grobHeight(title),pos=0)
panel6 <- gtable_add_grob(panel6,title,1, 1, 1, ncol(panel6))
rm(title)
```

The five panels and table are defined in the following *chunk*:

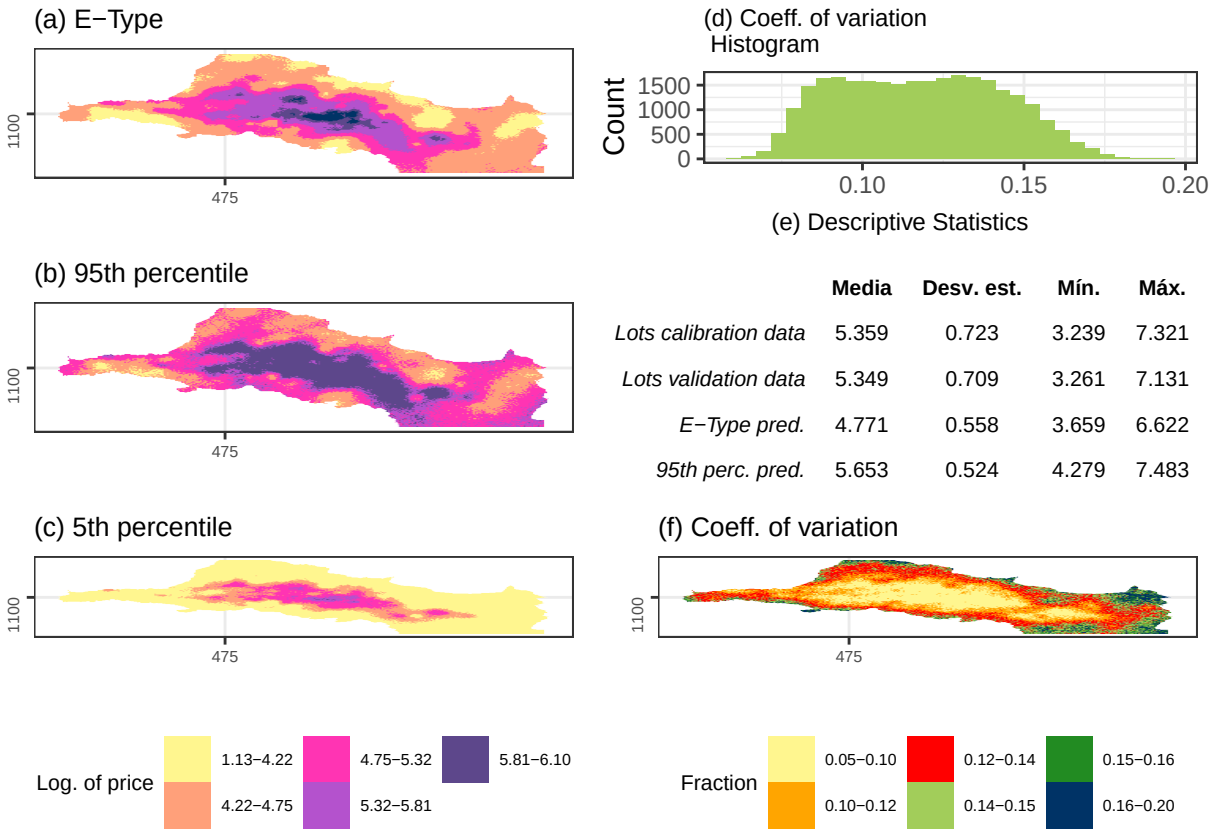
```
sim.full <- sim.full %>%
  mutate(long = unlist(map(sim.full$geometry,1)),
         lat = unlist(map(sim.full$geometry,2)))
# The E-type prediction
panel1 <- ggplot() +
  geom_raster(data = sim.full , aes(x=long/1000,y=lat/1000,fill = as.factor(CatMean)))+
  scale_fill_manual(values=c("khaki1","lightsalmon","maroon1",
                             "mediumorchid3","mediumpurple4","#003366"),
                    labels=c("1.13-4.22","4.22-4.75","4.75-5.32",
                             "5.32-5.81","5.81-6.10","6.10-8.31")) +
  labs(title="(a) E-Type",x="",y="") +
  scale_y_continuous(breaks = c(1100)) +
  scale_x_continuous(breaks = c(475)) +
  theme_bw() +
  theme(legend.position = "none",text = element_text(size = 8),
        axis.text.y = element_text(angle = 90))
# The 95th percentile
panel2 <- ggplot() +
  geom_raster(data = sim.full , aes(x=long/1000,y=lat/1000,fill = as.factor(Catq95)))+
  scale_fill_manual(values=c("khaki1","lightsalmon","maroon1",
                             "mediumorchid3","mediumpurple4","#003366"),
                    labels=c("1.13-4.22","4.22-4.75","4.75-5.32",
                             "5.32-5.81","5.81-6.10","6.10-8.31")) +
  labs(title="(b) 95th percentile",x="",y="") +
  scale_y_continuous(breaks = c(1100)) +
  scale_x_continuous(breaks = c(475)) +
  theme_bw() +
  theme(legend.position = "none",text = element_text(size = 8),
        axis.text.y = element_text(angle = 90))
# The 5th percentile
panel3 <- ggplot() +
  geom_raster(data = sim.full , aes(x=long/1000,y=lat/1000,fill = as.factor(Catq05)))+
  scale_fill_manual(values=c("khaki1","lightsalmon","maroon1",
                             "mediumorchid3","mediumpurple4","#003366"),
                    labels=c("1.13-4.22","4.22-4.75","4.75-5.32",
                             "5.32-5.81","5.81-6.10","6.10-8.31")) +
  labs(title="(c) 5th percentile",x="",y="",fill="Log. of price") +
  scale_y_continuous(breaks = c(1100)) +
  scale_x_continuous(breaks = c(475)) +
  theme_bw() +
  theme(legend.position = "bottom",text = element_text(size = 8),
        axis.text.y = element_text(angle = 90)) +
  guides(fill = guide_legend(nrow = 2))

# CV map
panel4 <- ggplot() +
  geom_raster(data = sim.full , aes(x=long/1000,y=lat/1000,fill = as.factor(Catcv)))+
  scale_fill_manual(values=c("khaki1","orange","red","darkolivegreen3",
                             "forestgreen","#003366"),
                    labels=c("0.05-0.10", "0.10-0.12", "0.12-0.14",
                             "0.14-0.15", "0.15-0.16","0.16-0.20")) +
  labs(title="(f) Coeff. of variation",fill="Fraction",x="",y="") +
```

```
  scale_y_continuous(breaks = c(1100)) +
  scale_x_continuous(breaks = c(475)) +
  theme_bw() +
  theme(legend.position = "bottom",text = element_text(size = 8),
        axis.text.y = element_text(angle = 90)) +
  guides(fill = guide_legend(nrow = 2))
# CV histogram
panel5 <- ggplot() +
  geom_histogram(data=as.data.frame(sim.full),
                 aes(x=cv),fill="darkolivegreen3") +
  ggtitle("(d) Coeff. of variation \n Histogram") +
  labs(y="Count",x="") + theme_bw() +
  theme(plot.title = element_text(size=9))
```

And we now plot the result:

```
m <- matrix(c(1,2,1,2,1,2,1,2,1,2,
              3,4,3,4,3,4,3,4,3,4,
              5,6,5,6,5,6,5,6,5,6,5,6,5,6), ncol=2, byrow=TRUE)
#jpeg(filename="G:/20230815/16_ValorSueloIncertidumbre/03_Reporte/Figuras/Fig02_maps.jpg",
#     width=6.6,height=7,units="in",res=600)
#jpeg(filename="D:/TEMPforCALC/16_ValorSueloIncertidumbre/03_Reporte/Figuras/Fig02_maps.jpg",
#     width=6.6,height=7,units="in",res=600)
grid.arrange(panel1,panel5,
             panel2,panel6,
             panel3,panel4,layout_matrix = m)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



	Media	Desv. est.	Mín.	Máx.
Lots calibration data	5.359	0.723	3.239	7.321
Lots validation data	5.349	0.709	3.261	7.131
E−Type pred.	4.771	0.558	3.659	6.622
95th perc. pred.	5.653	0.524	4.279	7.483

```
#dev.off()
```

Explain uncertainty patterns

We will now explore the impact of determinants on the location-wise estimated variance. To do so, we will subdivided the locations according to a given determinant and we will compare the statistical distribution among these groups. Should statistically significant differences be found, we will then compare the ecdf's of the groups in order to determine the relative magnitude and direction of the influence of the factor.

First, we create a data set with the variance estimate and the corresponding determinants (slope, distance to CBD, distance to municipal centers, elevation, distance to main roads):

```
det.analysis <- as.data.frame(sim.full)
det.analysis <- det.analysis %>% select(sd,geometry)
det.temp <- as.data.frame(st_grid)
sum((det.analysis$geometry==det.temp$geometry))#We check they perfectly coincide; they do because the s
```

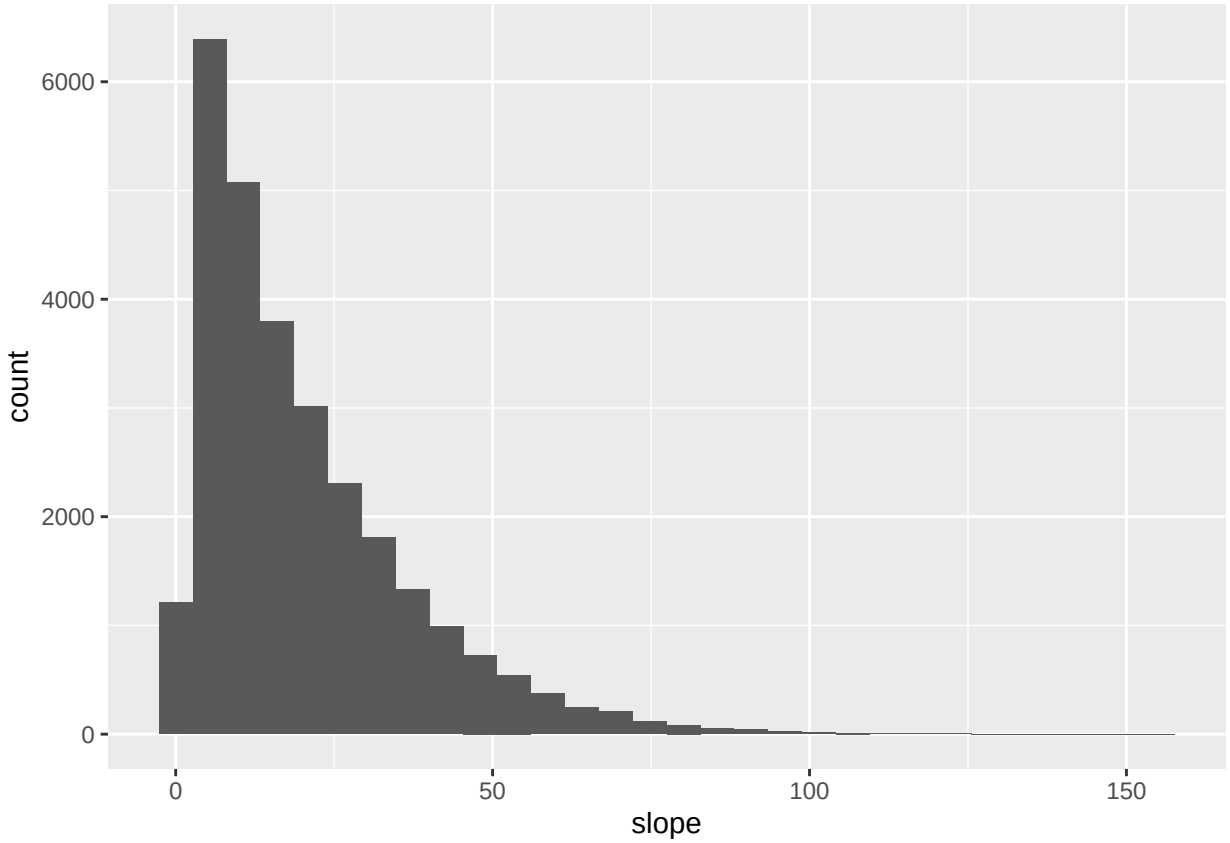
```
## [1] 28357
```

```
det.analysis <- cbind(det.analysis,det.temp)
det.analysis <- as.data.frame(det.analysis)
rm(det.temp)
det.analysis <- det.analysis %>% select(-geometry)
```

We examine the histograms of the determinants and create groupings:

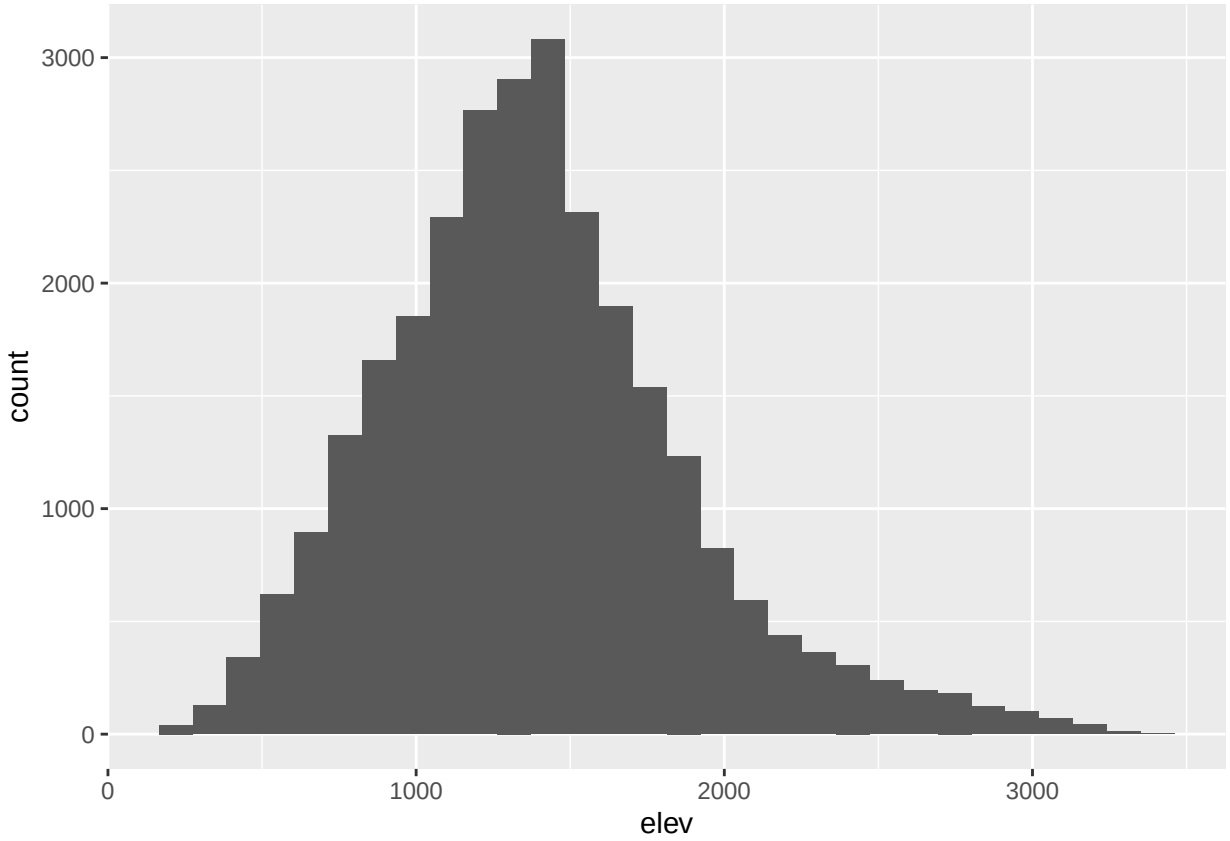
```
ggplot() + geom_histogram(data=det.analysis,aes(x=slope))#We use 30 and 50 from law
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



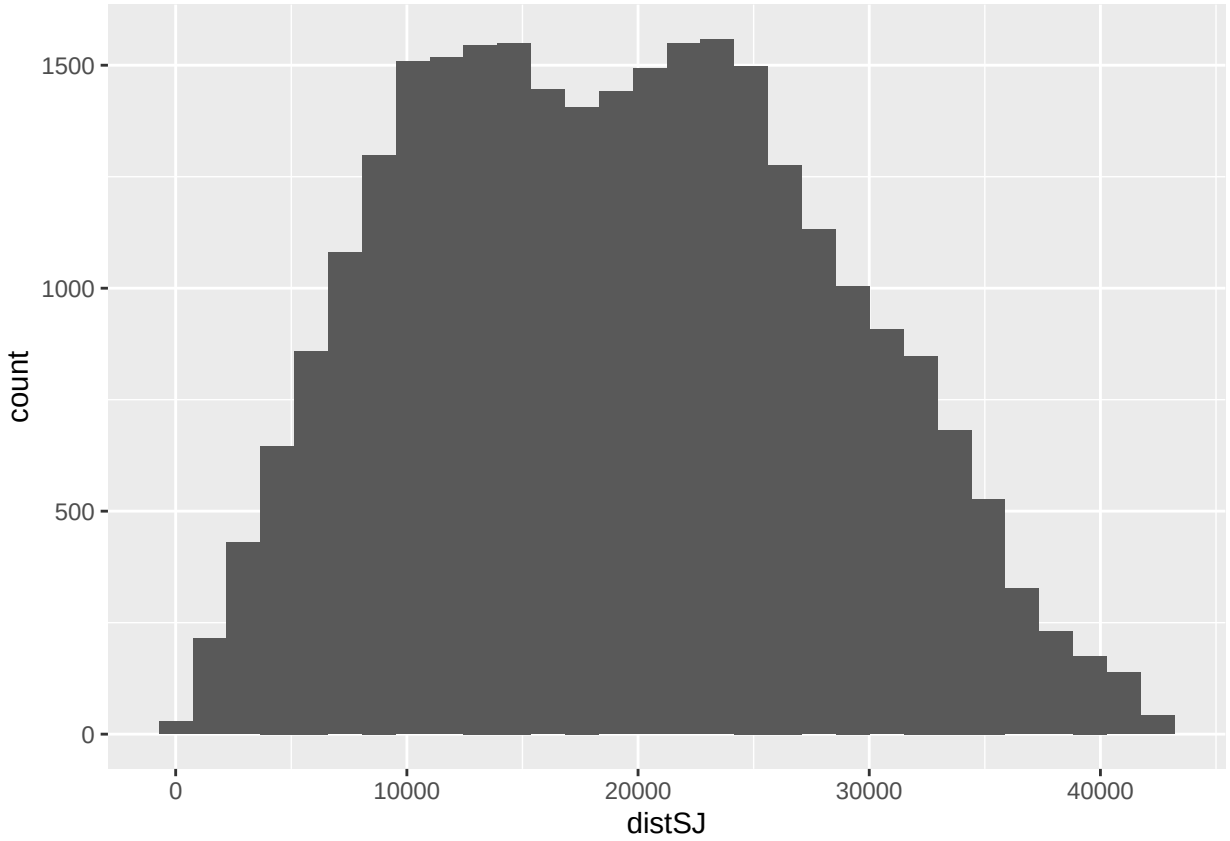

```
ggplot() + geom_histogram(data=det.analysis,aes(x=elev))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



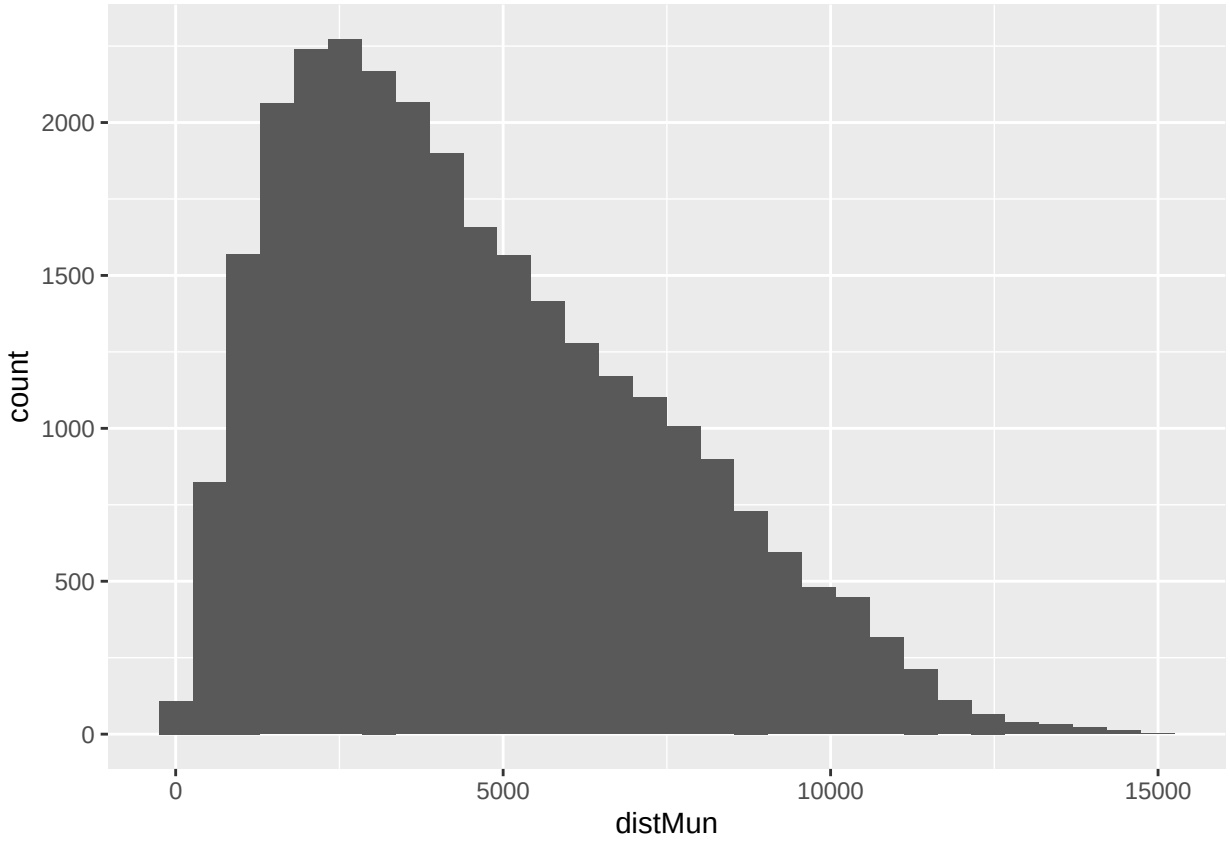

```
#Nothing especially salient, 1500 is the peak
ggplot() + geom_histogram(data=det.analysis,aes(x=distSJ))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



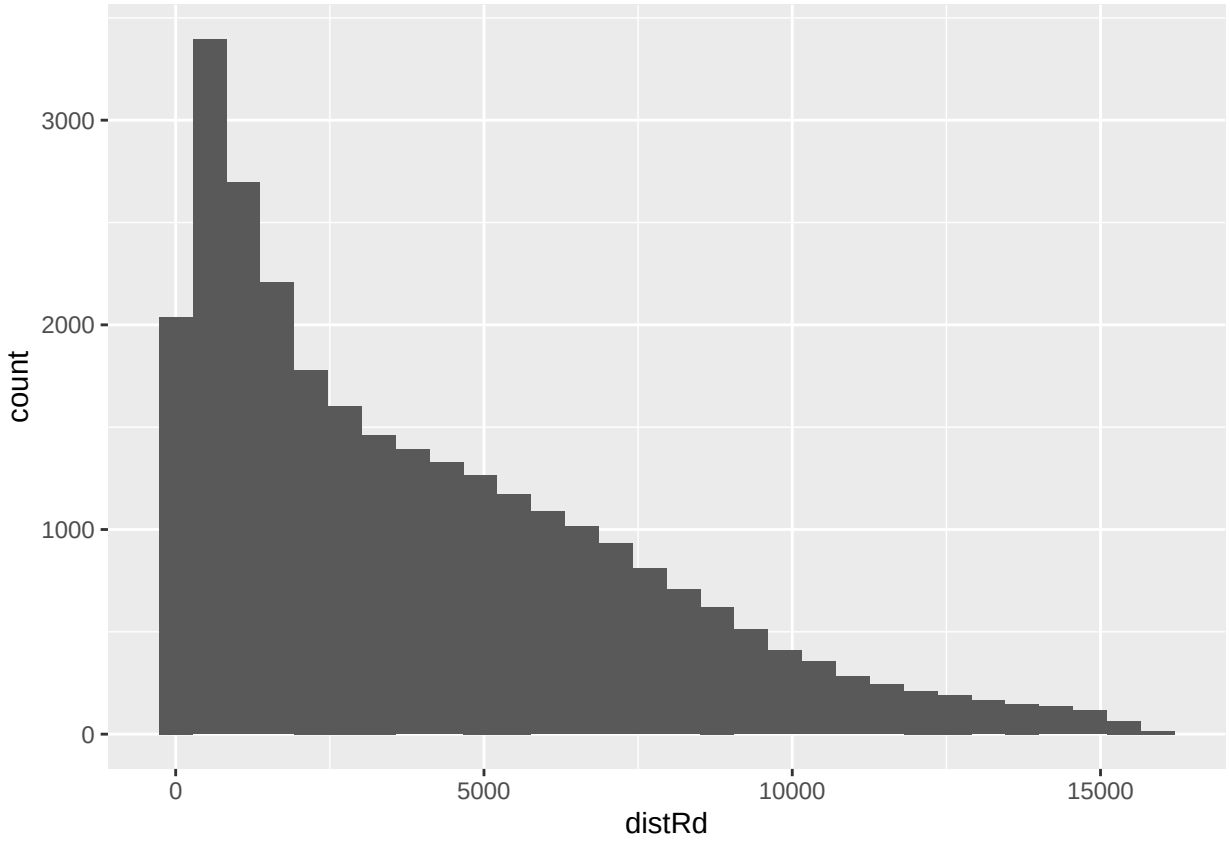

```
#less than 10k, more than 25k are the tails
ggplot() + geom_histogram(data=det.analysis,aes(x=distMun))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#something around 2500; 7500 as a second (to divide somehow)
ggplot() + geom_histogram(data=det.analysis,aes(x=distRd))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#something around a 1000m; rest is smooth: we choose 7500
#(roughly half the remaining range)
```

```
#Slope:
det.analysis$CatSlope <- 1
det.analysis$CatSlope[det.analysis$slope>30] <- 2
det.analysis$CatSlope[det.analysis$slope>50] <- 3
#Elevation
det.analysis$CatElev <- 1
det.analysis$CatElev[det.analysis$elev>1500] <- 2
#Dist SJ
det.analysis$CatDSJ <- 1
det.analysis$CatDSJ[det.analysis$distSJ>10000] <- 2
det.analysis$CatDSJ[det.analysis$distSJ>25000] <- 3
#Dist Mun
det.analysis$CatDmun <- 1
det.analysis$CatDmun[det.analysis$distMun>2500] <- 2
det.analysis$CatDmun[det.analysis$distMun>7500] <- 3
#Dist Road
det.analysis$CatDrd <- 1
det.analysis$CatDrd[det.analysis$distRd>1000] <- 2
det.analysis$CatDrd[det.analysis$distRd>7500] <- 3
```

We then apply the Kolmogorov-Smirnov test

```
#Slope tests:
ks.test(det.analysis$sd[det.analysis$CatSlope==1],det.analysis$sd[det.analysis$CatSlope==2])
```

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  det.analysis$sd[det.analysis$CatSlope == 1] and det.analysis$sd[det.analysis$CatSlope == 2]
## D = 0.34743, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
ks.test(det.analysis$sd[det.analysis$CatSlope==1],det.analysis$sd[det.analysis$CatSlope==3])
```

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  det.analysis$sd[det.analysis$CatSlope == 1] and det.analysis$sd[det.analysis$CatSlope == 3]
## D = 0.47409, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
ks.test(det.analysis$sd[det.analysis$CatSlope==2],det.analysis$sd[det.analysis$CatSlope==3])
```

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  det.analysis$sd[det.analysis$CatSlope == 2] and det.analysis$sd[det.analysis$CatSlope == 3]
## D = 0.1405, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
#Elevation test:
ks.test(det.analysis$sd[det.analysis$CatElev==1],det.analysis$sd[det.analysis$CatElev==2])
```

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  det.analysis$sd[det.analysis$CatElev == 1] and det.analysis$sd[det.analysis$CatElev == 2]
## D = 0.43611, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
#Distance to SJ test:
ks.test(det.analysis$sd[det.analysis$CatDSJ==1],det.analysis$sd[det.analysis$CatDSJ==2])
```

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  det.analysis$sd[det.analysis$CatDSJ == 1] and det.analysis$sd[det.analysis$CatDSJ == 2]
## D = 0.34869, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
ks.test(det.analysis$sd[det.analysis$CatDSJ==1],det.analysis$sd[det.analysis$CatDSJ==3])
```

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  det.analysis$sd[det.analysis$CatDSJ == 1] and det.analysis$sd[det.analysis$CatDSJ == 3]
## D = 0.60063, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
ks.test(det.analysis$sd[det.analysis$CatDSJ==2],det.analysis$sd[det.analysis$CatDSJ==3])
```

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
```

```
##
## data:  det.analysis$sd[det.analysis$CatDSJ == 2] and det.analysis$sd[det.analysis$CatDSJ == 3]
## D = 0.30724, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
#Distance to Mun test:
ks.test(det.analysis$sd[det.analysis$CatDmun==1],det.analysis$sd[det.analysis$CatDmun==2])
```

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  det.analysis$sd[det.analysis$CatDmun == 1] and det.analysis$sd[det.analysis$CatDmun == 2]
## D = 0.37775, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
ks.test(det.analysis$sd[det.analysis$CatDmun==1],det.analysis$sd[det.analysis$CatDmun==3])
```

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  det.analysis$sd[det.analysis$CatDmun == 1] and det.analysis$sd[det.analysis$CatDmun == 3]
## D = 0.70641, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
ks.test(det.analysis$sd[det.analysis$CatDmun==2],det.analysis$sd[det.analysis$CatDmun==3])
```

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  det.analysis$sd[det.analysis$CatDmun == 2] and det.analysis$sd[det.analysis$CatDmun == 3]
## D = 0.4013, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
#Distance to Rd test:
ks.test(det.analysis$sd[det.analysis$CatDrd==1],det.analysis$sd[det.analysis$CatDrd==2])
```

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  det.analysis$sd[det.analysis$CatDrd == 1] and det.analysis$sd[det.analysis$CatDrd == 2]
## D = 0.26871, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
ks.test(det.analysis$sd[det.analysis$CatDrd==1],det.analysis$sd[det.analysis$CatDrd==3])
```

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  det.analysis$sd[det.analysis$CatDrd == 1] and det.analysis$sd[det.analysis$CatDrd == 3]
## D = 0.62426, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
ks.test(det.analysis$sd[det.analysis$CatDrd==2],det.analysis$sd[det.analysis$CatDrd==3])
```

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  det.analysis$sd[det.analysis$CatDrd == 2] and det.analysis$sd[det.analysis$CatDrd == 3]
```

```
## D = 0.36695, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

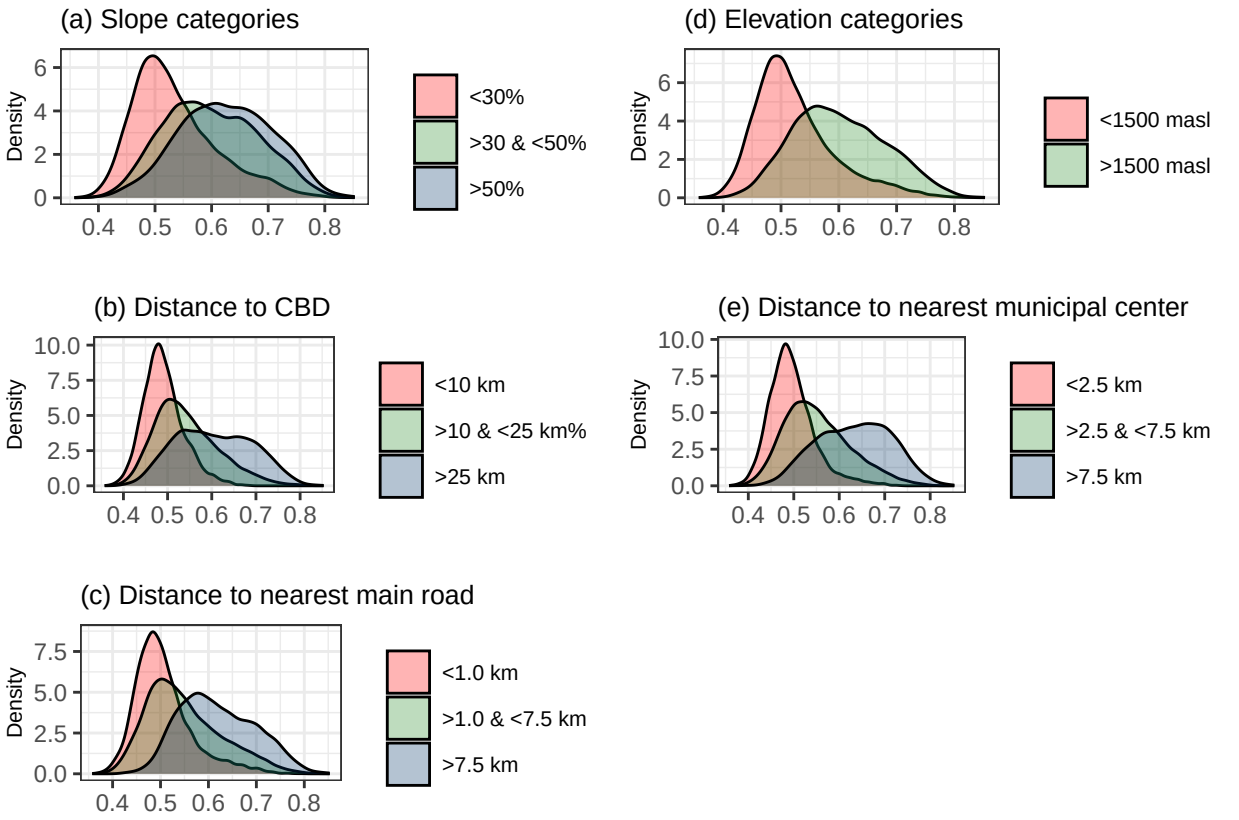
We create the panels with the ecdfs for the determinants with expected difference:

```
# Density plots with semi-transparent fill
panel7 <- ggplot(det.analysis, aes(x=sd, fill=as.factor(CatSlope))) +
  geom_density(alpha=.3) +
    scale_fill_manual(values=c("red","forestgreen","#003366"),
                      labels=c("<30%",">30 & <50%",">50%")) +
  ggtitle("(a) Slope categories") +
  labs(y="Density",x="",fill="") + theme_bw() +
  theme(legend.position="right",
                     legend.text=element_text(size=8),
                     legend.title=element_text(size=8),
                     axis.title=element_text(size=8),
                     plot.title = element_text(size=9.5))
panel8 <- ggplot(det.analysis, aes(x=sd, fill=as.factor(CatElev))) +
  geom_density(alpha=.3) +
    scale_fill_manual(values=c("red","forestgreen"),
                      labels=c("<1500 masl",">1500 masl")) +
  ggtitle("(d) Elevation categories") +
  labs(y="Density",x="",fill="") + theme_bw() +
  theme(legend.position="right",
                     legend.text=element_text(size=8),
                     legend.title=element_text(size=8),
                     axis.title=element_text(size=8),
                     plot.title = element_text(size=9.5))
panel9 <- ggplot(det.analysis, aes(x=sd, fill=as.factor(CatDSJ))) +
  geom_density(alpha=.3) +
    scale_fill_manual(values=c("red","forestgreen","#003366"),
                      labels=c("<10 km",">10 & <25 km%",">25 km")) +
  ggtitle("(b) Distance to CBD") +
  labs(y="Density",x="",fill="") + theme_bw() +
  theme(legend.position="right",
                     legend.text=element_text(size=8),
                     legend.title=element_text(size=8),
                     axis.title=element_text(size=8),
                     plot.title = element_text(size=9.5))
panel10 <- ggplot(det.analysis, aes(x=sd, fill=as.factor(CatDmun))) +
  geom_density(alpha=.3) +
    scale_fill_manual(values=c("red","forestgreen","#003366"),
                      labels=c("<2.5 km",">2.5 & <7.5 km",">7.5 km")) +
  ggtitle("(e) Distance to nearest municipal center") +
  labs(y="Density",x="",fill="") + theme_bw() +
  theme(legend.position="right",
        legend.text=element_text(size=8),
        legend.title=element_text(size=8),
        axis.title=element_text(size=8),
        plot.title = element_text(size=9.5))
panel11 <- ggplot(det.analysis, aes(x=sd, fill=as.factor(CatDrd))) +
  geom_density(alpha=.3) +
    scale_fill_manual(values=c("red","forestgreen","#003366"),
                      labels=c("<1.0 km",">1.0 & <7.5 km",">7.5 km")) +
  ggtitle("(c) Distance to nearest main road") +
```

```
  labs(y="Density",x="",fill="") + theme_bw() +
  theme(legend.position="right",
                      legend.text=element_text(size=8),
                      legend.title=element_text(size=8),
                      axis.title=element_text(size=8),
                      plot.title = element_text(size=9.5))
```

And finally we plot the result for reporting:

```
#jpeg(filename="F:/20230815/16_ValorSueloIncertidumbre/03_Reporte/Figuras/Fig03_ecdfs.jpg",
#     width=6.6,height=6,units="in",res=600)
grid.arrange(panel7,panel8,
             panel9,panel10,
             panel11,ncol=2)
```




```
#dev.off()
```