# An Application of Finite Mixture Models in Positional Accuracy

*Positional accuracy has always been considered a defining and essential element of the quality of any geospatial data, as it affects factors such as geometry, topology, and thematic quality; and it is directly related to the interoperability of spatial data. For its assessment, many procedures have been developed and many of them (for instance, EMAS or NMAS tests) and many of them require the underlying hypothesis of normality, which, however, is not easily found in sample data, but can be adequately adjusted by means of a finite mixture of normal distributions. In this work an application of the finite mixture model is presented for data from an Airborne Laser Scanner (ALS) campaign flight over the province of Ávila (Spain). In this case altimetric errors have been obtained on different types of terrain. These data can be adequately modelled by a gaussian finite mixture model through an adequate estimation procedure, and the subsequent parametric model is obtained. This model fits accurately the real data set and that be employed in further inferential analysis better than the assumption of normal distribution.*

**Keywords:** *Finite mixture models, ALS data, ML estimation, positional accuracy, parametric model*

## Introduction

The hypothesis of normality in the case of error measurements appears from the beginning of the normal distribution itself, since Laplace and Gauss arrived at it by analyzing measurement errors in astronomical observations. The fact that some errors or residuals are normally distributed implies that they are due to pure chance, and there are no other causes that explain them. In addition, the normality hypothesis is basic when it comes to proposing the contrasts of hypotheses on errors, both for mean values and for variances. However, in practice it is difficult to find measurement error data that are distributed according to a normal distribution.

One possible reason is that, although the data is normally distributed, it comes from different normal distributions. Therefore, one way of analysis is to assume that the data come from a finite mixture of distributions. In fact, in general, finite distribution models of distributions provide a mathematical basis for the resolution of multiple random phenomena, that is, they work with a tool equivalent to what in signal analysis consists of decomposing a signal by means of a series of sine/cosine functions (Fourier transform). They also allow the approximation for very complex distributions and allow to solve situations in which a single parametric distribution cannot provide a satisfactory model for local variations in the observed data. (McLachlan-Peel, 2000). The first works date back to 1894 when Pearson worked with the mixture of two normal distributions with the same variance and it has been developed by multiple researchers (a detailed review can be seen in McLachlan-Peel, 2000;

McLachlan et al 2019, or Huang et al, 2017 and some examples of recent applications of mixtures in different fields can be seen in Pan et al, 2020; Liu et al, 2020; Sallay et al, 2021; Zhao et al 2021; Li et al, 2021; Rodríguez-Avi, 2022).

One field in which the normality of errors plays a very important role is in the quality of cartographic data, since most standards have normality as their underlying hypothesis (Ariza-López and Atkinson-Gordo, 2008). Specifically, in digital elevation models (DEM) quality is understood as the positional accuracy of the data in its altimetric component. DEMs are topographic data that following a model (for instance, contour lines, point clouds, meshes, triangle networks, etc.) digitally represent the elevations (elevations or altimetry) of the terrain naked. The DEMs are data of great relevance and have been included as a theme of INSPIRE and the UN. DEMs have application in numerous branches of science and engineering and are mainly used to calculate the height, slope, orientation and delimitation of basins (Ariza-López et al. 2018).

There are multiple procedures proposed for the evaluation of positional accuracy (Mesa-Mingorance and Ariza-López 2020), although the most usual way is to apply standardized methods (Ariza-López et al, 2018), among which the National Map Accuracy Standards, NMAS (USBB 1947)), Engineering Map Accuracy Standard, EMAS (ASCE 1983) or the National Standard for Spatial Data Accuracy, NSSDA (FGDC 1998), the ASPRS proposals (ASPRS 1990, 2015) or the EuroSDR proposal based on measurements with a parametric approach (Höhle and Potuckova 2011). However, there are multiple studies that show that the normality hypothesis cannot always be accepted as true (Zandbergen 2008, 2011, Maune 2007), which has led to the development of other procedures in which this hypothesis is not necessary. (Höhle and Höhle, 2009; Ariza-López et al 2019; Cheok et al, 2008; Zandbergen, 2011).

Recently, the use of finite Gaussian mixtures has been proposed as a statistical procedure to propose a parametric model in the distribution of errors in DEM (Rodríguez-Avi and Ariza-López, 2022). The underlying assumption is that these data really come from another distribution, but in many cases, it is due to the fact that the set of observations has been obtained from different normal distributions, with different means and/or variances, and the mixture of all these data in a single data set does not have to follow a single normal distribution. In these situations, the data show possible multimodality, and/or a more sharped shape than the well-known Gaussian bell and the finite mixture technique of Gaussian distributions can be applied. It consists of decomposing the data into multiple normal distributions, estimating the mean and the corresponding variance, as well as the probability in the mixture. In this way, a population model can be generated that allows a better understanding of the nature of the analyzed variable.

In this work we propose the use of this procedure to model a set of altimetric errors to evaluate a lidar flight in relation to higher accuracy observations taken on the ground. Thus, the following section presents a summary of the statistical methodology used. Subsequently, the data are

presented, the best model is selected and the accuracy of the theoretical fit with the observed data is checked, ending with a discussion of the main results.

**Finite Mixture Models.**

This paper proposes an approach based on the use of Gaussian finite mixture models. In this context, the aim is to determine, through their parameters, which are the normal distributions that are mixed in the observed data set.

In a theoretical point of view, it is assumed that the vector of observed errors $X = (X_1, \dots, X_n)$ is a random sample that come from a mixture of $g > 1$ arbitrary distributions of probability. Then, the density function of each $X_i$ is given by:

$$f_\theta(x_i) = \sum_{i=1}^{g} \pi_j \phi_j(x_i); \quad x_i \in \mathbb{R} \qquad (1)$$

where $\boldsymbol{\Theta} = (\boldsymbol{\pi}, \boldsymbol{\phi}) = (\pi_1, \dots, \pi_g, \phi_1, \dots, \phi_g)$ is the vector of parameters in such a way that $\pi_1 + \cdots + \pi_g = 1$, $\pi_i > 0 \; \forall i$, and $(\phi_1, \dots, \phi_g)$ is the vector of parameters of each mixing distribution that comes from any absolutely continuous probability distribution family, $\mathcal{F}$. In our case it is considered that $\mathcal{F} = \{\phi(\cdot|\, \mu, \sigma)\}$ is the family of density functions $\mathcal{N}(\mu, \sigma)$, $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+$. In consequence, it is needed to estimate the vector $\boldsymbol{\Theta}$ of dimension $3g$ (eq. 2): the proportion of each density in the mixture ($\boldsymbol{\pi}$, a vector of dimension $g$) and the values of means and standard deviations for each distribution

$$\boldsymbol{\Theta} = (\pi_1, \dots, \pi_g, (\mu_1, \sigma_1), \dots, (\mu_g, \sigma_g)) \qquad (2)$$

In order to estimate (2) the EM algorithm (Dempster et al, 1977, Cueva-López et al, 2019) is applied, that provide an iterative solution of the calculus of Maximum Likelihood Estimators (MLE) in problems with missing values. The use of the EM algorithm is suggested not only for evidently incomplete data (missing values, truncated distributions, censored or grouped distributions), but also for statistical models where the absence of data is not so evident (McLachlan – Krishnan, 2008, McLachlan et al, 2019) as occurs with distributions obtained as mixtures. This algorithm uses, in an iterative way, the operator:

$$\mathrm{Q}\left(\theta \middle| \theta^{(\mathrm{t})}\right) = \mathrm{E}\left[\log h_\theta(C)\,|x, \theta^{(t)}\right] \qquad (3)$$

where $\theta \in \Theta, \theta^{(t)}$ is the value obtained at iteration t and the expectation refers to the distribution of $k_\theta(c|x)$ of c given x for the value $\theta^{(t)}$ of the parameter. Each iteration has two steps: (i) E-step where $\mathrm{Q}\left(\theta \middle| \theta^{(\mathrm{t})}\right)$ is computed and (ii) M-step where these values are used to maximize the likelihood of the mixing distribution and obtain the updated estimates $\theta^{(t+1)}$

Once parameters have been estimated, and by the Bayes theorem, it proceeds to make a probabilistic grouping to assign each value of the original set (or of the whole population), to the corresponding normal distribution to which has more pertaining probability, according to the posterior probabilities $\hat{\pi}_{ij}$ that $x_j$ belongs to the group with density function $f_i$:

$$\hat{\pi}_{ij} = \frac{\hat{\pi}_i \, f_i(x_j|(\hat{\mu}_i, \hat{\sigma}_i))}{\sum_{k=1}^{g} \hat{\pi}_k \, f_k(x_j|(\hat{\mu}_k, \hat{\sigma}_k))} \tag{4}$$

where $x_j \in \mathbb{R},\ i = 1 \ldots g,\ j = 1, \ldots, n$

In this way, given an observed value, it is assigned to the corresponding normal distribution where this probability is maximum.

Finally, it is possible calculate densities in the mixed model, adding all probabilities for each individual in the g obtained models:

$$f(x_j) = \sum_{i=1}^{g} \hat{\pi}_{ij}\ , \quad \mathrm{x_j} \in \mathbb{R}^r \tag{5}$$

where $\hat{\pi}_{ij}$ are obtained in (4).

The theoretical model provides a whole description about the population and all population probabilities and parameters can be calculated. In this case, the mean and variance of the model through the mixture, can be obtained as follows:

- Mean:

$$\hat{\mu} = \sum_{i=1}^{g} \hat{\pi}_i \hat{\mu}_i \tag{6}$$

- Variance:

$$\hat{\sigma}^2 = \sum_{i=1}^{g} \hat{\pi}_i \hat{\sigma}_i^2 + \sum_{i=1}^{g} \hat{\pi}_i (\hat{\mu}_i - \hat{\mu})^2 \tag{7}$$

- Standard deviation

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} \tag{8}$$

## Results

*Data Description*

To apply the procedure described, a set of data corresponding to altimetric errors detected between two DEM models has been selected, one considered as the Product and the other taken as a reference as it is a source of greater accuracy. (A more detailed description of the data used can be seen in Ariza et al, 2019)

- Product (PROD): Elevation data obtained from a LiDAR flight campaign executed over the city of Ávila (Spain) in April 2012. A Leica ALS50_II sensor was used with a flight height of 1000 m, which resulted in an original ground spacing of about 2 points/$m^2$ over urban and open land areas. Additionally, and to validate the accuracy of the sensor, several test locations were used: A stretch of paved urban road (labeled Infrastructure), a block of flats with different heights (labeled Urban) and a rugged field area with rocks, hillsides and vegetation (labeled as natural).
- Reference (REF): reference data from a higher accuracy source was obtained using a mobile mapping system (MMS), Optech Lynx in May 2012.
- Altimetric discrepancies (Errors): In those points for which the reference data were obtained, the error in height was measured as $Error = PROD - REF$.

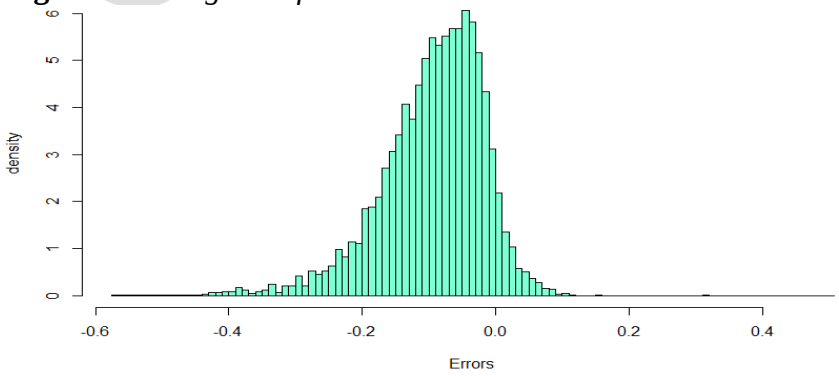
Once the errors were obtained, a total of N=5870 altimetric discrepancies were obtained between both sources. A descriptive analysis is shown in Table 1.

**Table 1**. *Valores descriptivos del conjunto de datos utilizado*

| Mean | Var | Sd | min | Q1 | Median | Q3 | max |
|---|---|---|---|---|---|---|---|
| -0.09446 | 0.00632 | 0.07950 | -0.57605 | -0.13611 | -0.08398 | -0.03957 | 0.50742 |

A graphical representation of the data appears on the histogram (Figure 1). It shows that the data shows some skewness to the left (the product data tends to underestimate the true height), as well as a sharp drop in errors greater than 0, which suggests the absence of normality in the set of errors.

**Figure 1.** *Histogram of Errors*



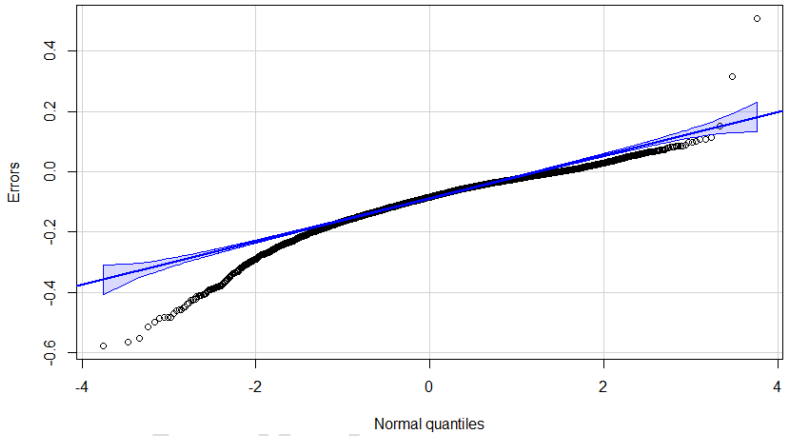
An analysis of the normality of the data through different contrasts is shown in Table 2. It can be seen how in the five contrasts proposed (asymmetry, kurtosis, Shapiro-Wilks, Kolmogorov-Smirnov and Jarque-Bera) the null hypothesis of normality in the data is rejected, as well as the presence of strong asymmetry to the left.

**Table 2.** *Normality tests*

| | *Coefficient* | *Test-statistic* | *p-value* |
|---|---|---|---|
| *Skewness* | -0.9846 | -30.7975 | 0 |
| *Curtosis* | 6.1308 | 95.9461 | 0 |
| *Shapiro-Wilks* | | 0.9277 | 0 |
| *Kolmogorov-Smirnov* | | 0.4694 | 0 |
| *Jarque-Bera* | | 10154.15 | 0 |

Graphically, the absence of normality can also be seen in the QQ-Plot (Figure 2), where it is observed how the cloud of points moves away from the diagonal line and the corresponding confidence interval (blue area).

**Figure 2.** *QQ-Plot*




*Model selection*

To obtain a fit of model (1), it is first necessary to fix the number of distributions in the mixture. To this end, its determination is proposed based on some of the usual criteria for comparing models, obtained from the value of the log-likelihood. In this case, the use of the following criteria is proposed (Anderson et al, 1998; Cameron-Trivedi, 2013; Burnham-Anderson, 2003):

- Akaike Information Criterion (AIC):

$$AIC = -2\mathcal{L} + 2p \qquad (9)$$

where $\mathcal{L}$ is the log-likelihood obtained from the estimation procedure and $p$ is the number of parameters, which, in this case, is $3g$, where $g$ is the number of distributions present in the mixture.

- Bayesian Information Criterion (BIC):

$$BIC = -2\mathcal{L} + p\,ln(N) \qquad (10)$$

- Consistent AIC (CAIC):

$$CAIC = -2\mathcal{L} + p\,(ln(N) + 1) \quad (11)$$

- Hannan-Quin Criterion (HQN):

$$HQN = -2\mathcal{L} + p(\ln(\ln(N)) \quad (12)$$

These measurements are related to the Kulblak Leibler distance and cannot be seen as hypothesis tests, that is, they are not valid to decide if a model is good or bad, but only serve to decide which is the best model, among several proposed ones. They consist of the log-likelihood plus a penalty term that benefits the parsimony of the model. The penalty term is related to the number of parameters so that more parameters (and therefore less parsimony), more penalty. The difference is that the AIC does not consider the size of the sample, while in the other three this value appears as a function of $\ln(N)$ to correct the tendency to overestimate that AIC presents. In any case, the best model will be the one for which the chosen criterion is minimal.

Consequently, to decide on the best model, the data has been adjusted by means of mixtures of $g$ normal distributions, (with $g = 2, \dots, 10$) and in each case the values of the aforementioned criteria have been calculated. The results are shown in Table 3, where it can be seen how for the AIC the best model is with 5 distributions (FMM5), while the other three criteria select 3 distributions (FMM3). In this paper, all calculations have been performed using the R package mixtools (R, 2022, Benaglia et al., 2008), which provides an estimate of the parameter vector $\boldsymbol{\Theta}$ given in equation (2).

**Table 3.** *Values of the different Information Criteria employed. In bold, the best model according to the corresponding column criterion*

| $g$ | $loglik$ | $AIC$ | $BIC$ | $CAIC$ | $HQN$ |
|---|---|---|---|---|---|
| 2 | 6937.70 | -13863.40 | -13823.34 | -13817.34 | -13849.47 |
| 3 | 7006.14 | -13994.28 | **-13934.18** | **-13925.18** | **-13973.39** |
| 4 | 7011.06 | -13998.12 | -13917.99 | -13905.99 | -13970.26 |
| 5 | 7016.04 | **-14002.08** | -13901.91 | -13886.91 | -13967.25 |
| 6 | 7017.62 | -13999.24 | -13879.04 | -13861.04 | -13957.45 |
| 7 | 7017.62 | -13993.25 | -13853.02 | -13832.02 | -13944.49 |
| 8 | 7021.25 | -13994.50 | -13834.23 | -13810.23 | -13938.78 |
| 9 | 7020.25 | -13986.50 | -13806.21 | -13779.21 | -13923.82 |
| 10 | 7020.25 | -13980.50 | -13780.17 | -13750.17 | -13910.86 |

We compare both models to see the differences and be able to decide. Tables 4 and 5 show the estimated parameters in each case, where $\mu$ are the means of each group, $\sigma$ the standard deviations and $\lambda$ the weight of each distribution in the mixture. In both cases, $\sum_{i=1}^{g} \lambda_i = 1$.

**Table 4.** *Set of estimated parameters, $FMM3$*

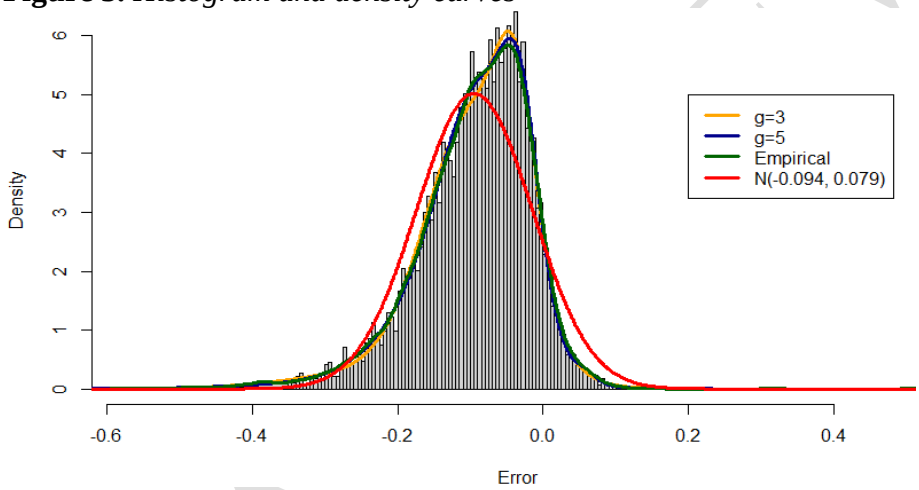
| Parameter | Group 1 | Group 2 | Group 3 |
|---|---|---|---|
| $\mu$ | -0.1877 | -0.0955 | -0.0364 |
| $\sigma$ | 0.1267 | 0.0633 | 0.0270 |
| $\pi$ | 0.1106 | 0.6989 | 0.1905 |

**Table 5.** *Set of estimated parameters, FMM5*

| | Group 1 | Group 2 | Group 3 | Group 4 | Group 5 |
|---|---|---|---|---|---|
| $\mu$ | -0.2031 | -0.1547 | -0.0816 | -0.0311 | 0.0589 |
| $\sigma$ | 0.1464 | 0.0708 | 0.0550 | 0.0244 | 0.0138 |
| $\pi$ | 0.0598 | 0.1883 | 0.5989 | 0.1484 | 0.0046 |

In consequence, according to (5) it is possible to determine the value of $f(x_j)$ for each model and also to compare this model with the empirical data. Figure 3 shows the histogram of the data together with the empirical density curve and the density curves of the models with 3 and 5 distributions in which we can see that they almost overlap and that the theoretical models adjust adequately to the observed data. Additionally, the density graph is shown if a univariate normal is assumed with the mean and standard deviation estimated from the data.

**Figure 3.** *Histogram and density curves*




Applying expressions (6) to (8), Table 6 shows the theorical mean (6), variance for each selected model, as well as the corresponding observed value. We observe that these parameters are perfectly well-fitted.

**Table 6.** *Mean, variance and standard deviation observed and fitted*

| Model | Mean | Variance | Standard dev. |
|---|---|---|---|
| Empirical | -0.09446 | 0.00632 | 0.07950 |
| $FMM3$ | -0.09446 | 0.00632 | 0.07950 |
| $FMM5$ | -0.09446 | 0.00632 | 0.07950 |

Additionally, it is possible to calculate the theorical distribution function for the two models, and, in consequence, to obtain probabilities for any interval. Table 7 shows calculated probabilities for different intervals, and we can compare these results with those provides by the data (empirical) and probabilities based on a normal distribution with the same mean and variance of the data, $\mathcal{N}(-0.09446, 0.07950)$

**Table 7.** *Comparison between distributions functions and the empirical distribution function*

| Value | DFE | FMM3 | FMM5 | Normal |
|---|---|---|---|---|
| -0.45 | 0.00238 | 0.00212 | 0.00274 | 0.00000 |
| -0.40 | 0.00511 | 0.00519 | 0.00539 | 0.00006 |
| -0.30 | 0.01856 | 0.02120 | 0.01899 | 0.00486 |
| -0.25 | 0.04003 | 0.03958 | 0.03983 | 0.02520 |
| -0.20 | 0.08688 | 0.08556 | 0.08894 | 0.09217 |
| -0.15 | 0.20306 | 0.20433 | 0.20136 | 0.24240 |
| -0.10 | 0.41073 | 0.41508 | 0.41344 | 0.47221 |
| -0.05 | 0.68739 | 0.68777 | 0.68796 | 0.71199 |
| 0.00 | 0.93236 | 0.92955 | 0.93138 | 0.88260 |
| 0.05 | 0.98909 | 0.98900 | 0.98865 | 0.96539 |
| 0.10 | 0.99880 | 0.99801 | 0.99852 | 0.99277 |
| 0.15 | 0.99948 | 0.99953 | 0.99951 | 0.99894 |

Similarly, Table 8 shows the comparison between the quartiles of the models compared to those of the Data distribution. It can also be seen how the mixture models adequately reproduce the observed quartiles, which does not occur when the adjustment is made using a normal distribution.

**Table 8.** *Comparison of Quartiles between models and the empirical distribution function*

| Quartil | DFE | FMM3 | FMM5 | Normal |
|---|---|---|---|---|
| Q1 | -0.13611 | -0.13739 | -0.13618 | -0.14808 |
| Median | -0.08398 | -0.08306 | -0.08339 | -0.09446 |
| Q3 | -0.03956 | -0.03971 | -0.03971 | -0.04083 |

Once both models are compared, it is observed that the model suggested as more adequate by the AIC criterion overestimates the number of distributions that make up the mixture, without a substantial improvement being observed with respect to the adjustment provided by the model suggested by the rest of the criteria, consisting of the mixture of three distributions, so it is the one that has been selected in this case.

*Group membership*

Since the theoretical model is obtained from several normal distributions, this process allows each of the points to be classified into a group, which corresponds to each distribution. To do this, and from expression (4), for each point it is possible to calculate the probability that it belongs to each distribution, and it is assigned to that group in which the probability is maximum.

From the selected theoretical model (the one made up of three normal distributions) each point can be assigned to the group to which it most likely belongs. That is, each point $x_i$ is assigned to the group where the value $\hat{\pi}_{ij}$ is

maximum. The result is shown in Table 9, where, let us remember, the points that belong to group 1 have a mean value of -0.188 m and a standard deviation of 0.127 m (that is, larger errors), group 2, to which the largest number of observations can be ascribed has a mean of -0.096 m and a standard deviation of 0.063, while the observations belonging to Group 3 have a mean of -0.036 m and a standard deviation of 0.027 m.

**Table 9.** *Number of points by group*

| Group | 1 | 2 | 3 |
|---|---|---|---|
| Number of points | 276 | 4886 | 708 |

*Group analysis*

The fact of obtaining groups without the need for additional information allows obtaining a new variable, qualitative in this case, which can be used to relate it to other variables, which may allow explaining the relationship of these groups through other additional variables, which can be both quantitative and qualitative. In this case, as has been mentioned in the description of the database, additional information is available on the type of terrain to which each point belongs.

Table 10 shows the descriptive analysis of the errors according to the type of terrain to which it belongs. It can be seen how there are differences in the means and standard deviations depending on the type of terrain:

**Table 10.** *Number of points by terrain type*

| Terrain type | N | Mean | Standard dev. |
|---|---|---|---|
| Abrupt | 4432 | -0,1084 | 0,0768 |
| Building | 863 | -0,0712 | 0,0829 |
| Road | 575 | -0,0217 | 0,0343 |

One of the advantages of the mixture model is that it allows obtaining a grouping variable, as has been obtained previously. In this way, it is possible to study whether there is a relationship between the intrinsic grouping (provided by the model) and the extrinsic grouping (obtained from the variable "terrain type"). Since both are qualitative, one way to study it is through the Contingency Table shown in Table 11.

**Table 11.** *Contingency table between "Terrain type" and "Group"*

| | | Group | | | Total |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | |
| Terrain type | Abrupt | 252 | 3801 | 379 | 4432 |
| | Building | 20 | 764 | 79 | 863 |
| | Road | 4 | 321 | 250 | 575 |
| Total | | 276 | 4886 | 708 | 5870 |

Based on Table 11, the independence tests are carried out, the results of which are shown in Table 12. In both cases, it is observed how the hypothesis of independence between the grouping provided by the model and the type of

terrain can be rejected.

**Table 12.** *Independence tests*

| | Statistics | df | p-value |
|---|---|---|---|
| Pearson's $\chi^2$ | 619.253 | 4 | .000 |
| likelihood ratio | 454.047 | 4 | .000 |

A first analysis of the relationship between both variables is shown in Table 13, where the standardized residuals are shown. Since the way to obtain them is from the difference between the observed and expected frequencies, a positive value indicates that there are more points than there should be if there were independence, and negative the opposite.

**Table 13.** *Standardized residuals*

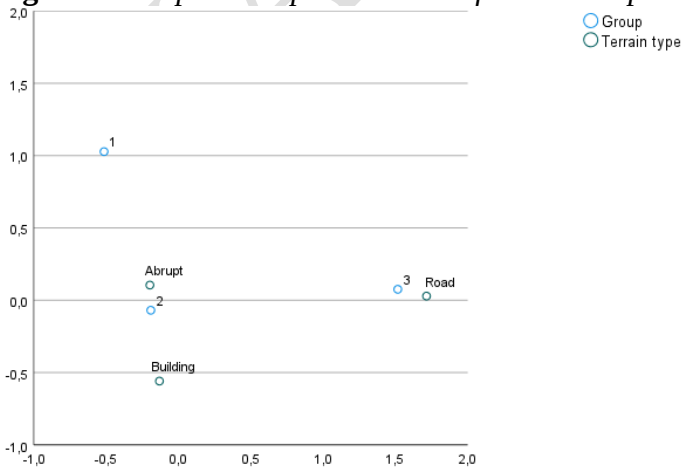
| | Group | | |
|---|---|---|---|
| Terrain type | 1 | 2 | 3 |
| Abrupt | 3,0 | 1,8 | -6,7 |
| Building | -3,2 | 1,7 | -2,5 |
| Road | -4,4 | -7,2 | 21,7 |

This relationship is confirmed by a Factorial Correspondence Analysis, as shown in Table 14 and Figure 4.

**Table 14.** *Summary of the Correspondence Analysis*

| Dimension | Singular value | Inertia | $\chi^2$ | Sig. | Inertia proportion | | singular value of confidence | |
|---|---|---|---|---|---|---|---|---|
| | | | | | accounted for | Cumulative | Standard dev. | Correlation 2 |
| 1 | .320 | .103 | | | .972 | .972 | .018 | .020 |
| 2 | .054 | .003 | | | .028 | 1.000 | .010 | |
| Total | | .105 | 619.253 | .000 | 1.000 | 1.000 | | |

**Figure 4.** *Graphical representation of the Correspondence Analysis*



## Discussion

Given a set of sample data, an important problem is to find a suitable parametric model that can describe and model the observed data, so that the extension of the results to the entire population may be possible. This is applicable to the case of errors in digital elevation models, especially when making decisions about the quality of the observations obtained, which are frequently made under the assumption that the errors are distributed according to a normal distribution. This assumption is correct from a theoretical point of view, since, if true, it implies that the observed errors are purely random and independent of each other. However, in most cases, these hypotheses of randomness and independence are not met, either for external reasons (for example, bias in the measurement instrument or errors in the data collection process) or internal reasons (for example, the fact that not all areas are homogeneous: a forest area is not the same as a desert area when it comes to accurately determining a height above sea level). These facts can mean that, when combining all the observations, the final result is not normally distributed.

For this reason, this paper presents an example of the application of the finite mixtures of distributions procedure -specifically Gaussian- to obtain a parametric model that can be proposed to model the observed data. It is a more complex model, since the parameters are the means and standard deviations of each of the mixing distributions, as well as the weighting of each distribution in the mixture, but once obtained, it provides us with a theoretical way to calculate the distribution function that best fits the data. We have commented that mixtures of Gaussian distributions are proposed here, but in theory other mixture distributions can be assumed, such as log normal, Weibull or even, in the case of errors taken in absolute value, half normal or skew-normal distributions.

The proposed working method consists of first determining the number of distributions that make up the mixture, g by using information criteria -in this case the BIC better than the AIC due to the principle of parsimony-, the determination of the 3g parameters, and the construction of the theoretical distribution function. In this way, we have the desired parametric model, which allows us to obtain probabilities for any event in the population. In Rodríguez-Avi and Ariza-López (2022) the model is used to propose hypothesis contrasts by studying the distributions of statistics associated with sampling in such a population, which leads to a rethinking of the usual quality standards. , such as the NMAS, EMAS or NSSDA, while here it is proposed to use the information about the model to divide the errors into groups, which allows them to be related to other available variables. Given the available data, there is only information on the type of terrain to which the measured point belongs, but this analysis can be more interesting if more additional information is available, such as slope, orientation, land use, etc. These analyzes can allow a better analysis of the causes of the errors, and, where possible, suggest procedures for their improvement.

We believe that this data analysis technique, in this case altimetry errors, can be applied in many other situations in which it is necessary to circumvent the limitations of non-normality that have been pointed out in multiple studies and in different situations.

## References


ASCE (1983). *Map Uses, Scales and Accuracies for Engineering and Associated Purposes.* American Society of Civil Engineers, Committee on Cartographic Surveying, Surveying and Mapping Division: New York, NY, USA, 1983.

Zandbergen PA. (2008) Positional Accuracy of Spatial Data: Non-Normal Distributions and a Critique of the National Standard for Spatial Data Accuracy. *Transactions in GIS,* 12, pp. 103-130. DOI:10.1111/j.1467-9671.2008.01088.x

Anderson DR, Burnham KP, White GC. (1998). Comparison of Akaike information criterion and consistent Akaike information criterion for model selection and statistical inference from capture-recapture studies. *Journal of Applied Statistics*, 25(2), 263-282

Ariza-López FJ, García-Balboa, JL, Rodríguez-Avi, J, Robledo J, (2018). *Guía general para la evaluación de la exactitud posicional de datos espaciales*. Propuesta de adopción de metodologías y procedimientos empleados para la evaluación de la calidad de la información geográfica para los Estados Miembros del IPGH (Proyectos Panamericanos de Asistencia Técnica – 2018 "Agenda del IPGH 2010-2020"). Montevideo. Accesible en: http://publicaciones.ipgh.org/publicaciones-ocasionales/Guia_Evaluacion_Exactitud_Posicional_Datos_Espaciales.pdf

Ariza-López FJ, Atkinson AD (2008). Analysis of Some Positional Accuracy Assessment Methodologies. *Surveying Engineering*, 134(2), pp. 404, 407. https://doi.org/10.1061/(ASCE)0733-9453(2008)134:2(45)

Ariza-López FJ, Chicaiza-Mora EG, Mesa-Mingorance JL, Cai J, Reinoso-Gordo JF (2018). DEMs: An Approach to Users and Uses from the Quality Perspective. *International Journal of Spatial Data Infrastructures Research*, 2018, Vol.13, 131-171. Special Section: INSPIRE (Full Research Article).

Ariza-López FJ, Mozas-Calvache, AT (2012) Comparison of four line-based positional assessment methods by means of synthetic data. *Geoinformatica* 16, 221–243. https://doi.org/10.1007/s10707-011-0130-y

Ariza-López, FJ, Rodríguez-Avi J, González-Aguilera, D, Rodríguez-Gonzálvez, P. (2019). "A New Method for Positional Accuracy Control for Non-Normal Errors Applied to Airborne Laser Scanner Data" *Applied Sciences* 9(18):3887. https://doi.org/10.3390/app9183887

ASCE (1983). *Map Uses, scales and accuracies for engineering and associated purposes*. American Society of Civil Engineers, Committee on Cartographic Surveying, Surveying and Mapping Division, New York, USA.

ASPRS (1990). Accuracy standards for large scale maps. *Photogrammetric Engineering and Remote Sciences*, 56(7), 1068-1070.

ASPRS (2015). ASPRS Positional accuracy standards for digital geospatial

data, P*hotogrammetric Engineering & Remote Sensing*, 81(3), 53 p. http://www.asprs.org/a/society/divisions/pad/Accuracy/Draft_ASPRS_Accuracy_Standards_for_Digital_Geospatial_Data_PE&RS.pdf

Benaglia T, Chauveau D, Hunter DR, Young D, (2009). mixtools: An R Package for Analyzing Finite Mixture Models. J*ournal of Statistical Software*, 32(6), 1-29. DOI 10.18637/jss. v032.i06

Burnham KP, Anderson DR. (2003). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer Science & Business Media: New York, NY, USA.

Cameron AC and Trivedi PK. (2013). *Regression Analysis of Count Data. Second edition*. New York, NY: Cambridge University Press.

Cheok G, Filliben J, Lytle AM (2008). *NISTIR 7638. Guidelines for Accepting 2D Building Plans*. National Institute of Standards and Technology: Gaithersburg, MD, USA.

Cueva-López V, Olmo-Jiménez MJ. Rodríguez-Avi J. (2019). EM algorithm for an extension of the Waring distribution. Computational and Mathematics Methods 1, e1046. https://doi.org/10.1002/cmm4.1046

Dempster A, Laird N. and Rubin D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39(1): 1-38.

FGDC (1998). FGDC-STD-007: *Geospatial Positioning Accuracy Standards, Part 3. National Standard for Spatial Data Accuracy*. Federal Geographic Data Committee, Reston, USA. https://www.fgdc.gov/standards/projects/accuracy/part3/chapter3

Höhle J, Höhle M. Accuracy assessment of digital elevation models by means of robust statistical methods. *ISPRS Journal of Photogrammetry and Remote Sensing*. 2009, 64, 398–406.

Höhle J, Potuckova M. (2011). *Assessment of the quality of DTM*. EuroSDR.

Huang T, Peng H and Zhang K. (2017) Model Selection for Gaussian Mixture Models. *Statistica Sinica* 27, 147-169. DOI 10.5705/ss.2014.105

Karlis D. EM algorithm for mixed Poisson and other discrete distributions. *ASTIN Bulletin*. 2005;35(1):3-24.

Li J, Du G, Clouser JM, Stromberg A, Mays G, Sorra J, Brock J, Davis T, Mitchell S, Nguyen HQ and Williams MV. (2021). Improving evidence-based grouping of transitional care strategies in hospital implementation using statistical tools and expert review. *BMC Health Services Research* 21:35. DOI: 10.1186/s12913-020-06020-9

Liu G, Lee CC, Liu Y. (2020) Growth path heterogeneity across provincial economies in China: The role of geography versus institutions. *Empirical Economics*, 2020, 59, 503–546. DOI: 10.1007/s00181-019-01639-y

McLachlan GJ, and Peel D. (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics, New York.

McLachlan GJ, Krishnan T. *The EM Algorithm and Extensions*. 2nd ed. Hoboken, NJ: John Wiley and Sons, Inc; 2008.

McLachlan GJ, Lee SX and Rathnayake SI. (2019). Finite Mixture Models. *Annual Review of Statistics and Its Application* 6:355–78. DOI: 10.1146/annurev-statistics-031017-100325

Maune, DF (Editor) (2007) *Digital Elevation Model Technologies and Applications: The Dem User's Manual*. American Society for Photogrammetry and Remote Sensing, Bethesda, ISBN 978-1-57083-082-2.

Mesa-Mingorance JL, Ariza-López FJ. (2020). Accuracy Assessment of Digital Elevation Models (DEMs): A Critical Review of Practices of the Past Three Decades. *Remote Sensing*. 2020; 12(16):2630. https://doi.org/10.3390/rs12162630

Pan Y, Xie L, Su H and Luo L. (2020). A Robust Infinite Gaussian Mixture Model and its Application in Fault Detection on Nonlinear Multimode Processes. *Journal of Chemical Engineering of Japan*, 53(12), 758-770. DOI : 10.1252/jcej.17we373

Polidori L, El Hage M. (2020). Digital Elevation Model Quality Assessment Methods: A Critical Review. *Remote Sensing*. 2020; 12(21):3522. https://doi.org/10.3390/rs12213522

R Core Team (2022). R: *A language and environment for statistical computing. R* Foun-dation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rodríguez-Avi J. (2022) A Probabilistic Model for the Distribution of GDP per Capita in NUTS 3 Zones of Europe. Studies on Applied Economy 40, 5326.

Rodríguez-Avi J.; Ariza-López FJ. (2022). Finite Mixture Models in the Evaluation of Positional Accuracy of Geospatial Data. *Remote Sensing*, 14, 2062. https://doi.org/10.3390/rs14092062

Sallay H, Bourouis S, Bouguila N. (2021). Online Learning of Finite and Infinite Gamma Mix-ture Models for COVID-19 Detection in Medical Image. *Computers* 10, 6. DOI: 10.3390/computers10010006

USBB (1947). *United States National Map Accuracy Standards*. U.S. Bureau of the Budget. Washington, USA.

Zandbergen PA. (2008). Positional Accuracy of Spatial Data: Non-Normal Distributions and a Critique of the National Standard for Spatial Data Accuracy. *Transactions in GIS* 12(1):103–130. https://doi.org/10.1111/j.1467-9671.2008.01088.x

Zandbergen PA. (2011). Characterizing the error distribution of Lidar elevation data for North Carolina. *International Journal of Remote Sensing* 32(2):409-430. https://doi.org/10.1080/01431160903474939

Zhao B, Yang F, Zhang R, Shen J, Pilz J and Zhang D. (2021) Application of unsupervised learning of finite mixture models in ASTER VNIR data-driven land use classification, Journal of Spatial Science, 66:1, 89-112, DOI: 10.1080/14498596.2019.1570478.