International Cartographic Association
Association Cartographique Internationale

Finite mixtures of normal distributions in the study of the error in altimetry

José Rodríguez-Avi[a,]*, Francisco Javier Ariza-López[b]

[a] *Department of Statistics and Operational Research, University of Jaén, Jaén, Spain- jravi@ujaen.es*
[b] *Department of Cartographic Engineering, Geodesy and Photogrammetry, University of Jaén, Jaén, Spain – fjariza@ujaen.es*

* Corresponding author

Abstract: The modelling of the altimetric error is proposed by means of the mixture of normal distributions. This alternative allows to avoid the problems of lack of normality of the altimetric error and that have been indicated numerous times. The conceptual bases of the mixture of distributions are presented and its application is demonstrated with an applied example. In the example, the altimetric errors existing between a DEM with 5x5 m resolution and another DEM with 2x2 m resolution are modelled, which is considered as a reference. The application demonstrates the feasibility and power of analysis of the proposal made.

Keywords: DEM, Positional Accuracy, Data Quality, Altimetric Errors, Statistical Modelling

1. Introduction

Digital Elevation Models (DEM) are topographic data that following a model (e.g., contour lines, point clouds, meshes, triangle networks, etc.) digitally represent the elevations (elevations or altimetry) of the bare terrain. DEMs have application in numerous branches of science and engineering and are used mainly for the calculation of height, slope, orientation and delimitation of basins (Ariza-López et al., 2018b). So that, DEM are considered as base data, and in this way the United Nations (UN-GGIM, 2019) and the European Union (https://inspire.ec.europa.eu/Themes/Data-Specifications/2892) consider them as a theme in their lists of relevant geospatial data layers.

In the geomatic field, the quality of DEMs is usually understood as the altimetric positional accuracy of point data. Given a data product (PRO) and a reference of higher accuracy (REF), the altimetric discrepancies, or errors, (d = ZPRO-ZREF) are analyzed in a given set of points. There are very numerous approaches and methods developed to evaluate this positional accuracy (Mesa-Mingorance and Ariza-López, 2020). The best way to evaluate or control positional accuracy is by applying standardized methods, for example the new ASPRS standard (ASPRS, 2015), but there are many others (see a current guide of the most outstanding in Ariza-López et al., 2019). Until now, these standards are based on the assumption of the normality of errors (e.g., ASCE 1983, FGDC 1998, ASPRS 2015) which allows to the application of a parametric model: the normal distribution where the mean (μ) and the standard distribution (σ) are the position and scale parameters of the distribution. However, many studies (Zandbergen 2008, 2011; Maune, 2007) indicate that positional errors are not normally distributed. Regarding the work with methods based on the assumption of normality of the data, the non-normality of these can have several consequences depending on the degree of non-normality and the robustness of the applied method. In this case, non-normality violates a basic assumption of the method, and this violation is important from a strict perspective.

The normal distribution is a suitable distribution to represent real-valued random variables. Therefore, fully adequate to describe the altimetric discrepancies d. However, the abundance of references indicating the non-normality of altimetric-discrepancy data leads us to look for alternative statistical approaches, for example approaches are the use of robust statistics (Höhle and Höhle, 2009), the use of tolerances based on observed distributions (Ariza-López and Rodríguez-Avi, 2018a).

In this work, a new way is explored, which consists in assuming that the altimetric discrepancies do not really come from a single normal distribution and that, on the contrary, they are the result of the mixture of several normal distributions. This way is very powerful, and interesting, since it consists of decomposing the observed-error density function into a composition of a certain number of normal functions such that they adequately approximate it, that is, we work with a tool equivalent to what in analysis of signals consists of decomposing a signal by means of series of sine/cosine functions (Fourier transform).

The underlying idea is that the observed variable really comes from a mixture of data from distributions that follow the same model (the normal), but with different parameters (means and standard deviations). In this way, the probability of an observed value comes from the mixture of the probabilities that it comes from each of the distributions that make up the mixture. The first works date back to 1894 when Pearson worked with the mixture of two normal distributions with the same variance and 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

Advances in Cartography and GIScience of the International Cartographic Association, 3, 2021.
30th International Cartographic Conference (ICC 2021), 14–18 December 2021, Florence, Italy. This contribution underwent double-blind peer review based on the full paper. https://doi.org/10.5194/ica-adv-3-13-2021 | © Author(s) 2021. CC BY 4.0 License

applications of mixtures in different fields can be seen, for instance, in Zhao et al., 2021 or Li et al., 2021).

To the best of our knowledge, this approach has never been previously applied to the case of errors in DEM. Thus, the objective of this paper is to propose the use of this well-known general-statistical approach for analyzing and describing the altimetric positional accuracy, and that can be applied to any type of error data from DEM.

This document is organized in the following way; section 2 presents a basic conceptual approach to the mixture of normal distributions. In section 3 an application method is proposed and in section 4 the proposed method is applied step by step to the case of data from two DEM with different spatial resolutions (2x2 m and 5x5 m). Section 5 presents the discussion and finally, some general conclusions are included. (ICA):

2. Mixture of Normal distributions

The assumption of error's normality in measures appears from the same origin of the normal distribution. Indeed, Gauss obtained it by studying astronomical errors. In fact, error's normality implies that there are not any external cause of errors, but pure chance.

Nevertheless, in many cases, a distribution of measure errors, even when the underlying normality is adequate, can be overall non-normally distributed. This occurs when errors come from normal distribution but with different parameters. In this case, the mixture of data originating from different normal distributions cannot be adequately modelled by a single normal distribution.

In this paper we propose an approach of this problem about the distribution of errors based on the use of Gaussian finite mixture models. In this context, we try to determine, through their parameters, which are the normal distributions that are mixed in the observed data set.

In a theoretical point of view, we assume 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

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

Where $\mathbf{\Theta} = (\boldsymbol{\pi}, \boldsymbol{\theta}) = (\pi_1, \dots, \pi_g, \theta_1, \dots, \theta_g)$ is the vector of parameters in such a way that $\pi_1 + \cdots + \pi_g = 1$, with $\pi_i > 0\ \forall i$ and $(\theta_1, \dots, \theta_g)$ is the vector of parameters of each mixing distribution that comes from any absolutely continuous probability distribution family, $\mathcal{F}$. In our case we consider 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, we need to estimate the vector of dimension $3g$:

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

In order to estimate (2) we utilize the EM algorithm (Dempster et al, 1977), 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:

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

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 $\mathrm{Q}(\theta|\theta^{(t)})$ 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, we proceed 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:

$$\hat{\pi}_{ij} = \frac{\hat{\pi}_i\, f_i(x_j|(\hat{\mu}_i, \hat{\sigma}_i))}{\sum_{\mathrm{m}=1}^{g} \hat{\pi}_m\, f_m(x_j|(\hat{\mu}_m, \hat{\sigma}_m))}, \qquad (4)$$

where $x_j \in \mathbb{R}$, g is the number of mixing distributions and $\hat{\pi}_{ij}$ is the posterior probability that x_j belongs to the group with density function f_i. In this way, given an observed value, it is assigned to the corresponding normal distribution where this probability is maximum.

In addition, we can calculate probabilities in the final mixed models, adding all probabilities for each point in the g obtained models:

$$P[X = x_j] = \sum_{i=1}^{g} \hat{\pi}_{ij}, \quad \mathrm{x_j} \in \mathbb{R}^r \qquad (5)$$

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

3. Description of data.

To show an example of an application on real data, a study area around Allo (Navarra, Spain) is used by means of the following digital elevation models:

- DEM02. It is a gridded DEM (2x2 meter resolution), it was generated in 2017 and its primary data source is a Lidar survey (second coverage of the PNOA-LiDAR project https://pnoa.ign.es/estado-del-proyecto-lidar/segunda-cobertura). It is considered as the reference in this example.
- DEM05. It is a gridded DEM (5x5 meter resolution), it was generated in 2012 and its primary data source is a Lidar survey (first coverage of the PNOA-LiDAR project https://pnoa.ign.es/estado-del-proyecto-lidar/primera-cobertura).

Both DEM data sets come from the Instituto Geográfico Nacional (IGN, Spain, www.igne.es) and are freely available. In relation to the study area, Figure 1 shows a general vision. The area is 504 km2, and it has a varied relief, but not abrupt, with valleys of different widths, and areas with different degrees of undulation. The elevation is in the interval 316-1046 m, mean value of elevation is 468 m and the standard deviation of the elevation is 92.8 m.

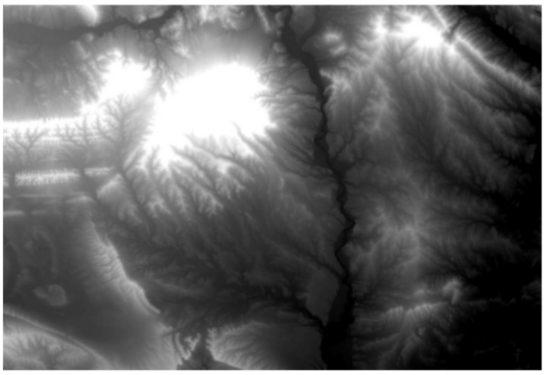

Figure 1: Study area (Allo, Navarra, Spain).

The subtraction of both models (DEM05 - DEM02) allows to obtain the altimetric discrepancy model. Assuming a global normal distribution for these discrepancies the relevant parameters values are: μ = -0.017 m, and σ = 0.280 m. Several experts took samples manually in the altimetric discrepancies model. These zones represent different altimetry discrepancy environments between DEM05 and DEM02. Table 1 presents the 15 labeled categories under consideration and the account of cases (DEM cells).

Code	Category	N
0	Tree-lined roadside	1045
10	Hilly	144353
20	Terraces	17159
30	Parcels' boundaries	3337
40	Built	44223
41	Built scattered	17922
50	Dense Forestry	13158
60	Water	3015
61	Paved roads	7380
62	Fallow land	19232
63	Plain	4246
64	Fruit trees	6133
65	Irrigated land	49826
66	Valley between hills	2507
70	Sloped scrub	5099
Total		338635

Table 1: Codes and Categories.

Additionally, Table 2 show a descriptive analysis of each category and the global data.

Code	Mean [m]	Median [m]	s.d. [m]	Min. [m]	Max [m]
0	-0.417	-0.463	0.318	-1.229	0.403
10	-0.113	-0.093	0.122	-2.517	1.732
20	0.351	0.403	0.457	-1.748	2.243
30	-0.140	-0.026	0.468	-2.054	1.258
40	0.024	0.038	0.307	-3.480	2.951
41	-0.190	-0.058	0.481	-2.788	2.477
50	0.198	0.112	0.437	-1.818	3.046
60	1.232	1.377	0.951	-0.691	3.094
61	-0.099	-0.101	0.116	-3.775	0.948
62	0.197	0.231	0.191	-0.228	0.601
63	-0.028	-0.005	0.078	-0.700	0.237
64	-0.129	-0.113	0.080	-0.568	0.179
65	-0.053	-0.062	0.084	-0.268	0.340
66	-0.119	-0.125	0.063	-0.317	0.073
70	0.129	0.122	0.205	-0.652	1.356
Total	-0.017	-0.060	0.280	-1.628	3.094

Table 2: Descriptive analysis.

Figure 2 shows the quantile-quantile plot and the expected normal distribution (blue line). In the three figures we observe the presence of several local modes, and a significate deviation from normality. That suggests the possibility of a mixture of normal distribution as an underlying data model.

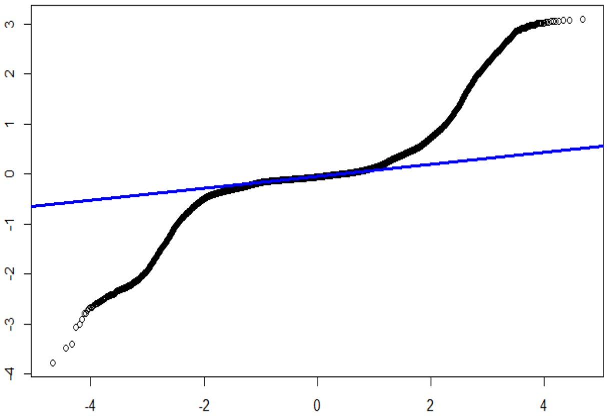

Figure 2: Normal Q-Q Plot.

Figure 3 shows the global histogram of the 338635 analyzed points, whereas Figure 4 shows this histogram restricted to the interval (-0.5, 0.5), where we can observe an irregular shape with some local modes.

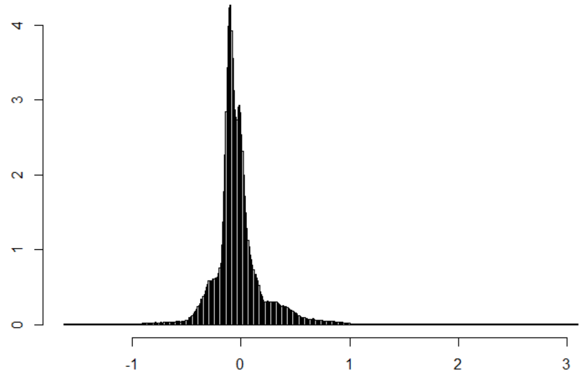

Figure 3: Histogram of altimetric discrepancies.

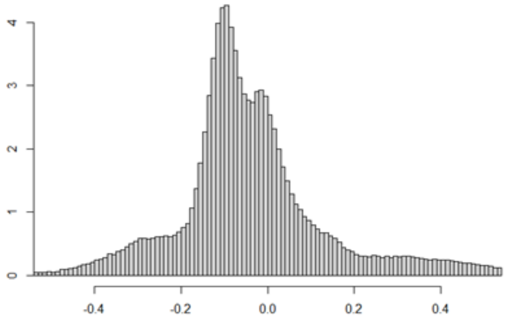

Figure 4: Histogram of altimetric errors (from -0.5 to 0.5).

4. Method: the mixture model selection

The first step is to determine g, the number of mixing normal distributions that compound the global mixture, to reproduce the observed form of errors. To deal with this problem we apply some information criteria to decide the best model (see for instance Cameron-Trivedi, 2013; Burnham-Anderson, 2003). Concretely we calculate the Akaike Information Criteria (AIC), defined as:

$$AIC = -2\mathcal{L} + 2p, \qquad (5)$$

where $\mathcal{L}$ is the value of the log-likelihood obtained through the estimation procedure, n is the sample size and p in the number of parameters, in this case, $3g$. We have to take into account that this measure, related to the Kulblak Leibler distance is not a contrast about the model goodness of fit, but only make a comparison between models, in the way that the best model is which the obtained value is minimum.

Table 3 shows AIC criteria when g comes from 2 to 9, and we observe that the minimum value is accomplish for $g = 8$. All calculations have been carried out using the package *mixtools* of R (R, 2021, Benaglia et al., 2008), that provide an estimation of the parameter vector $\mathbf{\Theta}$ given in (2). According to Table 3, we propose a mixture of 8 normal distributions with different parameters to model the distribution of altimetric errors.

g	AIC	g	AIC
2	-157032.2	6	-169964.5
3	-157026.2	7	-170064.7
4	-160257.4	8	-171074.8
5	-168505.7	9	-170100.0

Table 3: Values of AIC for g from 2 to 9.

Table 4 shows the vector of estimated parameters, $\widehat{\mathbf{\Theta}}$, obtained, where (μ_i, σ_i) are the parameters of the i normal distribution and π_i its probability

normal	μ_i	σ_i	π_i
1	0.3094	0.1401	0.0743
2	0.1184	0.5602	0.1171
3	-0.1238	0.0339	0.1503
4	-0.0968	0.0258	0.1463
5	-0.2497	0.0955	0.1313
6	-0.0191	0.0429	0.2765
7	0.086	0.062	0.1004
8	2.0172	0.4489	0.0039

Table 4: Estimated parameters when $g = 8$.

We can see that there is a small set of data that are extremely dispersed, and only 27% of the data correspond to errors with a mean of practically 0. It is also worth noting high values of the standard deviations and that the probabilities of belonging are very distributed among the groups (an exception to the last, which are the most positively biased). Figure 5 graphically represents the fit of the 8 normal distributions (colored lines) and the fit to the true observed density, which is the dotted line.

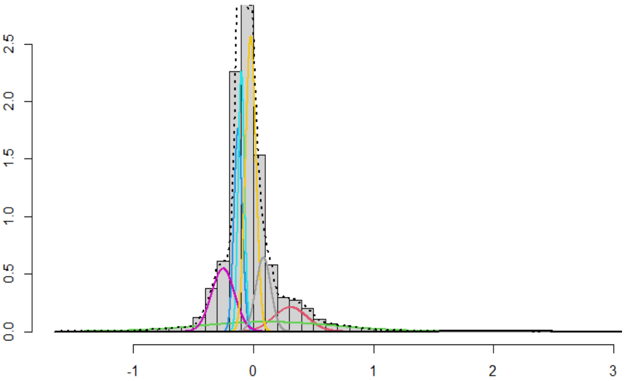

Figure 5: Graphical representation with 8 normal.

Once the number of mixing distribution is determined and parameters are estimated, the next step is about the behaviour of the resulting mixed distribution. For this, and using (4) we analyse all sampling points to the distribution to which it is most likely to belong. In this case, Table 5 shows the number of sampling points that can be assigned to each group, where the groups are those that appear in Table 4.

g	Number of points
1	28575
2	17263
3	49076
4	62715
5	41490
6	104194
7	34068
8	1254

Table 5: Number of sampling points assigned to each group.

In this way, and using (5), we can obtain the mixed theoretical distribution, and we compare it with the observed empirical distribution. Figures 6 and 7 show a superposition of the histogram of the empirical distribution and the theoretical density curve (in orange) obtained from the estimated normal distributions. In Figure 7 case, only values from -1 to 1 are drawn because the high range of errors. Additionally, in this Figure we show the normal density (in green) with mean -0.017 m and standard deviation 0.280 m.

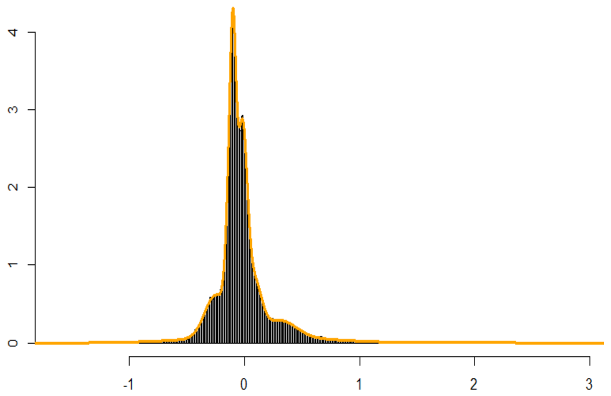

Figure 6: Observed Histogram and expected density of altimetric errors.

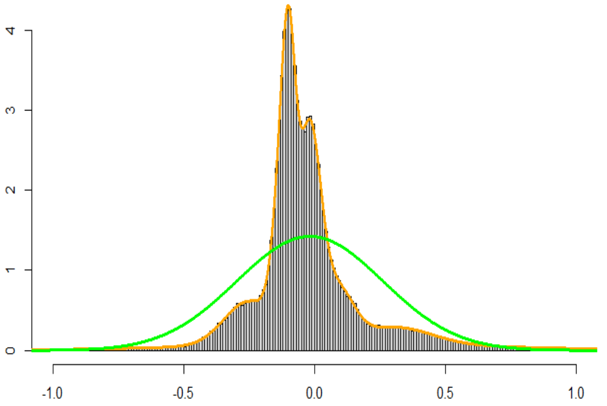

Figure 7: Observed Histogram and expected density of altimetric errors (from -1 to 1). In green, normal density considering that all data come to a single normal.

5. Results and analysis

Starting from the specified model, we can obtain probabilities using it and compare it with the empirical distribution and the one obtained assuming a normal $\mathcal{N}(-0.017, 0.280)$. Table 6 shows several probabilities obtained in these three ways:

- In the column labeled "Empirical model", values are the relative frequency of points that verify the row condition,
- In the column labeled "Mixture model", probabilities are obtained using (4). That is to say, is the weighted (by the value of $\hat{\pi}_i$) sum of the eight normal distribution functions
- In the column labeled "One Normal", probabilities are directly obtained from a Normal distribution with mean -0.017 m and standard deviation 0.280 m.

Value	Empirical model	Mixture Model	One Normal
$X > 1.1$	0.0081	0.0085	3.28e-5
$X > 0$	0.3285	0.3289	0.4751
$X < -0.125$	0.2523	0.2526	0.3505
$X > 1.5$	0.0042	0.0042	3.0e-08
$X < -1$	0.0027	0.0027	0.0002
$-0.1 < X < 0.1$	0.4723	0.4718	0.2785

Table 6: Comparison between empirical model, mixed model and the 1-Normal model.

In all cases we observe that the probabilities obtained through the mixed model are very close to the observed relative frequencies provided by the data, especially for the values in the tails.

Once the points have been assigned, in the form that each point x_j is assigned to the group where the value $\hat{\pi}_i\, f_i(x_j|(\hat{\mu}_i, \hat{\sigma}_i))$is maximum (Table 5), the next step consists on relating these new eight groups with the initial point classification according to its terrain type and that appears on Table 1. Due that both variables (ground type and assigned group) are qualitative, a contingency table relating the type of terrain and the distribution to which it has been assigned can be built. This contingency table appears in Tables 7 and 8, where the value n_{ij} indicates the number of points that comes from the terrain type i and have been assigned to the group j

Terrain	*a posteriori* Group			
type	N1	N2	N3	N4
0	15	525	58	34
10	223	317	22111	38629
20	6908	6948	206	159
30	610	748	153	130
40	5483	2662	2713	3268
41	488	1975	1803	2235
50	3024	2366	785	622
60	118	1391	25	13
61	4	0	1618	4795
62	10092	201	194	1454
63	0	0	1072	161
64	0	0	1720	2211
65	62	0	15300	8187
66	0	0	1091	594
70	1548	130	227	223

Table 7: Number of sampling points by type that belongs to groups 1-8.

Terrain	*a posteriori* Group			
type	N5	N6	N7	N8
0	224	129	60	0
10	30296	49340	3437	0
20	824	635	1459	20
30	569	503	624	0
40	3707	12987	13403	0

41	2132	7557	1732	0
50	995	2391	2975	0
60	189	25	20	1234
61	86	575	302	0
62	6	4420	2865	0
63	13	2607	393	0
64	1146	1043	13	0
65	690	20445	5142	0
66	326	484	12	0
70	287	1053	1631	0

Table 8: Number of sampling points by type that belongs to groups 1-8 (continuation).

Starting from this contingency table we are interested in studying if both variables are related. In this case we can apply the Pearson's χ^2 test for independence in contingency tables, where the null hypothesis is that the variables have no relationship between them. In this case, the value χ^2 = 412294.4 that under the null hypothesis follows a χ^2 distribution with 98 degrees of freedom (d.o.f) so the p-value is 0. Similarly, the likelihood ratio test produces a 247898.4 statistic, which follows the same χ^2 distribution with 98 d.o.f and its p-value is also 0. In fact, if we calculate the Pearson contingency coefficient is:

$$C_p = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{412294}{412294 + 338635}} = 0.7410 \qquad (6)$$

The proportion of points of each type of terrain that belong to each group are given in Tables 9 and 10. The most frequent model for each terrain type are highlighted in bold.

Terrain	*a posteriori* Group			
type	N1	N2	N3	N4
0	1.44	**50.24**	5.55	3.25
10	0.15	0.22	15.32	26.76
20	**40.26**	**40.49**	1.20	0.93
30	18.28	22.42	4.58	3.90
40	12.40	6.02	6.13	7.39
41	2.72	11.02	10.06	12.47
50	22.98	17.98	5.97	4.73
60	3.91	**46.14**	0.83	0.43
61	0.05	0.00	21.92	**64.97**
62	**52.48**	1.05	1.01	7.56
63	0.00	0.00	25.25	3.79
64	0.00	0.00	28.05	36.05
65	0.12	0.00	30.71	16.43
66	0.00	0.00	**43.52**	23.69
70	30.36	2.55	4.45	4.37

Table 9: Proportion of sampling points by type that belongs to groups 1-8.

Terrain	*a posteriori* Group			
type	N5	N6	N7	N8
0	21.44	12.34	5.74	0.00
10	20.99	34.18	2.38	0.00
20	4.80	3.70	8.50	0.12
30	17.05	15.07	18.70	0.00
40	8.38	29.37	30.31	0.00
41	11.90	**42.17**	9.66	0.00
50	7.56	18.17	22.61	0.00
60	6.27	0.83	0.66	**40.93**
61	1.17	7.79	4.09	0.00
62	0.03	22.98	14.90	0.00
63	0.31	**61.40**	9.26	0.00
64	18.69	17.01	0.21	0.00
65	1.38	**41.03**	10.32	0.00
66	13.00	19.31	0.48	0.00
70	5.63	20.65	31.99	0.00

Table 10: proportion of sampling points by type that belongs to groups 1-8 (continuation).

It is noteworthy that 40% of the elements of land type 60 belong to distribution 8, which is the one with the highest average, which is called “Water”. Similarly, in distributions 3, 5 and 7 there are no groups for which they are dominant.

On the other hand, having the mixture model allows it to be applied to all the positions of the DEM and to obtain a spatial representation of the probability of belonging to each altimetric discrepancy value to each of the normal distributions that make up the mixture. This is what is presented in Figures 8 and 9 for groups N1 and N8, respectively. As can be seen, Figure 8 presents greater probabilities in the northern half, which is consistent with the terrain types where N1 participates. The distribution shown in Figure 9 is consistent with those areas, with a reduced dimension, where the highest values of altimetric discrepancy occur.

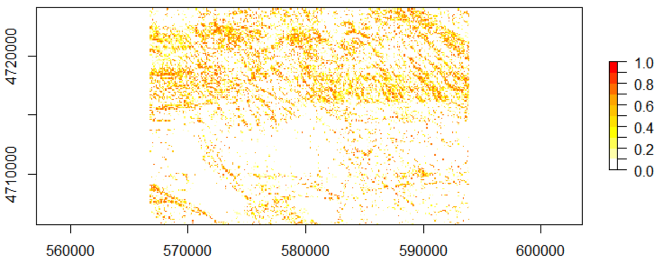

Figure 8: Spatial distribution of probabilities of Normal 1

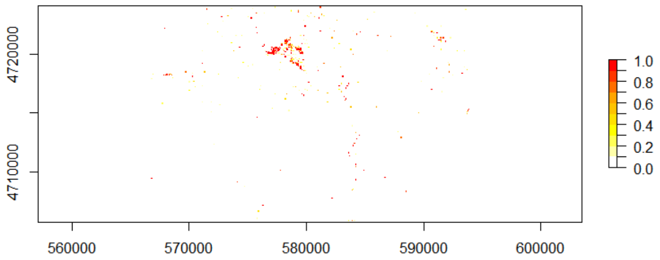

Figure 9: Spatial distribution of probabilities of Normal 8

6. Discussion

The hypothesis of normality underlying measurement errors presupposes the fact that these errors are random, independent and have no other external cause that influences them (pure chance). Nevertheless, in practice, in many cases the normality hypothesis is not fulfilled, as is the example here shown (see Fig.4). This non-normality can be attributed to multiple causes. One of these causes, that is very common, is the fact that even though measurement errors are normally distributed, they come from a mixture of several normal distributions with different parameters. In this case the overall population, obtained as an addition of data from these different normal distributions, is not, itself, a normal distribution. In the case that has been presented, it has been considered that the mixture of normal distributions come from different topographic and land cover situations, but it is also true that the parameters of the data collection by the LiDAR sensor (e.g., height, angle of incidence, humidity of the ground, etc.), influence altimetric errors.

The great discussion that can be raised about the previous approach is whether to work with or without prior information. If additional information is available on the points, the groups obtained by assigning each point to the mixing normal distribution to which it is most likely to belong can be related with that additional information. In our case, this information is qualitative -the type of terrain-, but it can be of any other type. This helps to better understand the cause of the error and identify, for example, those areas most prone to extreme errors. All of this information can be useful when designing more precise quality control procedures. If this information is not available, the method is also capable of determining the number of mixing normal distributions and their parameters; however, in this case it may be more complex to find an obvious physical meaning for the categories that are made up.

The examples of probabilities presented in Table 6 clearly indicate the goodness of fit to the observed data. In addition, it is also evident that the option of adjusting a single global normal is a bad approximation for this case. Tables 8 and 9 present results that show that there are types of terrain that are better defined than others, understanding as better defined that they present an explanation based on few groups, that is, a high proportion of cases in those few groups (e.g., terrain types 0, 61, 63). There are also other cases where this is not the case (e.g., type 30). It is interesting to note that group 8, which is responsible for the largest outliers in the general normal model, is concentrated in a single type of terrain (type 60). These results show us that, perhaps, the terrain types classification being considered, or the samples taken from them, are not the most appropriate, but it is not a problem to understand that the method of mixtures works properly.

The parameter estimation method also requires a comment. We have used the EM method to estimate the parameters that correspond to each of the distributions that make up the mixture. The EM is a well-known algorithm that is available in many computerized calculation tools and, as shown, its use is relatively straightforward. This algorithm is for general use in optimization and is included in packages suitable for determining the parameters of normal mixtures, as is the case of the *mixtools* package of R (R, 2021, Benaglia et al., 2008), that has been used in this work. If this is possible, as in the example presented above, we can reproduce the empirical data distribution function through a theoretical model, which, once obtained, can be extended to the entire population. In this way, we go from having a model that cannot be explained by a single normal, and that can be considered as non-parametric, to a parametric model with multiple parameters. This provides a new way of analysis since the methods based on a single normal can be applied to models based on mixtures of normal distributions. Of course, the last will be more complex, but not intractable.

7. Conclusions

An application of the finite mixture analysis techniques of distributions to the case of altimetric discrepancies in a DEM is presented. The theoretical basis and the tools for its application already exist, so it is not risky or expensive to apply them to this field.

A method for this type of analysis was developed, and a practical example has been carried out based on real data, that shows how to carry out this application, as well as the results obtained.

We consider that this technique for the analysis of discrepancies in altimetry allows us to apply the conventional positional quality assessment methods to this mixture model, which opens a line that extends its possibilities and avoids the limitations of non-normality that arise in many studies of error in altimetry. So that, looking ahead, an interesting research line is to analyze how the positional accuracy assessment standards can be applied when a characterization of the errors through mixtures is available.

This procedure may also be applied in many other cases involving continuous data where normality can be assumed in advance.

8. Acknowledgements

This work has been partially financed by the research project "Functional Quality in Digital Models of Land Elevations in Engineering" (https://coello.ujaen.es/investigacion/web_giic/funquality4dem/) of the State Agency research of Spain. PID2019-106195RB-I00 /AEI/10.13039/501100011033,

9. References

Ariza-López F.J., García-Balboa, J.L., Rodríguez-Avi, J., Robledo J., (2018). Guía general para la evaluación de la exactitud posicional de datos espaciales. Instituto Panamericano de Geografía e Historia, Méjico. http://publicaciones.ipgh.org/publicaciones-ocasionales/Guia_Evaluacion_Exactitud_Posicional_Datos_Espaciales.pdf

Ariza-López F.J., Rodríguez-Avi J., Alba-Fernández V. (2018a) A Positional Quality Control Test Based on Proportions. In: Mansourian A., Pilesjö P., Harrie L., van Lammeren R. (eds) Geospatial Technologies for All. AGILE 2018. Lecture Notes in Geoinformation and Cartography. Springer, Cham. https://doi.org/10.1007/978-3-319-78208-9_18

Ariza-López, F.J.; Chicaiza-Mora, E.G.; Mesa-Mingorance, J.L.; Cai, J.; Reinoso-Gordo, J.F. (2018b). DEMs: An Approach to Users and Uses from the Quality Perspective. Int. J. Spat. Data Infrastruct. Res. 2018, 13, 131–171. DOI: https://doi.org/10.2902/1725-0463.2018.13.art12

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 (2015). ASPRS Positional Accuracy standards for digital Geospatial Data. Photogrammetric Engineering & Remote Sensing, vol 81(4), pp.53-63, 2015

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

Burnham and Anderson (2003) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer Science & Business Media.

Cameron, A. C. and Trivedi, P.K. (2013). *Regression Analysis of Count Data*. Second edition. New York, NY: Cambridge University Press.

Dempster, A., Laird, N. y 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., M. Höhle M. (2009) "Accuracy assessment of digital elevation models by means of robust statistical methods". ISPRS J. Photogrammetry and Remote Sensing, vol. 64(4), pp 398–406. DOI: 10.1016/j.isprsjprs.2009.02.003

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

Li J , Du G, Clouser JM, Stromberg A, Mays G, Sorra J, Brock J, Davis T, Mitchell S, Ngu-yen HQ & 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

Maune, D.F. (2007). Digital Elevation Model. Technologies and Applications: The DEM Users Manual, 2nd ed.; American Society for Photogrammetry and Remote Sensing: Bethesda, MD, USA, 2007

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

McLachlan, G. J.; Lee, S. X. and Rathnayake, S. I. (2019) Finite Mixture Models. *Annu. Rev. Stat. Appl.* **6**:355–78. DOI: 10.1146/annurev-statistics-031017-100325

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

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

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

UN-GGIM (2019). The Global Fundamental Geospatial Data Themes. http://ggim.un.org/documents/Fundamental%20Data%20Publication.pdf

Zandbergen P. A. (2008). "Positional Accuracy of Spatial Data: Non-Normal Distributions and a Critique of the National Standard for Spatial Data Accuracy". Transactions in GIS, vol 12(1), pp.103–130. DOI: 10.1111/j.1467-9671.2008.01088.x

Zandbergen P. A. (2011). "Characterizing the error distribution of Lidar elevation data for North Carolina". *International Journal of Remote Sensing*, vol, 32(2), pp. 409-430. DOI: 10.1080/01431160903474939

Zhao B, Yang F, Zhang R, Shen J, Pilz J & 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.