Quality control method for non-normal positional error data using a multinomial approach

Francisco Javier Ariza-López, José Rodríguez-Avi ,Virtudes Alba-Fernández, José Luis García-Balboa

[fjariza, jravi, mvalba, jlbalboa]@ujaen.es; chamorro@us.es

Universidad de Jaén; Paraje de las Lagunillas S/N, E-23.071-Jaén

Funded by the Project: MEMCALIG (Modelización Estadística de Matrices de Confusión en Calidad de la Información Geográfica)

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

PROYECTO COFINANCIADO POR LA UNIÓN EUROPEA
FONDO EUROPEO DE DESARROLLO REGIONAL

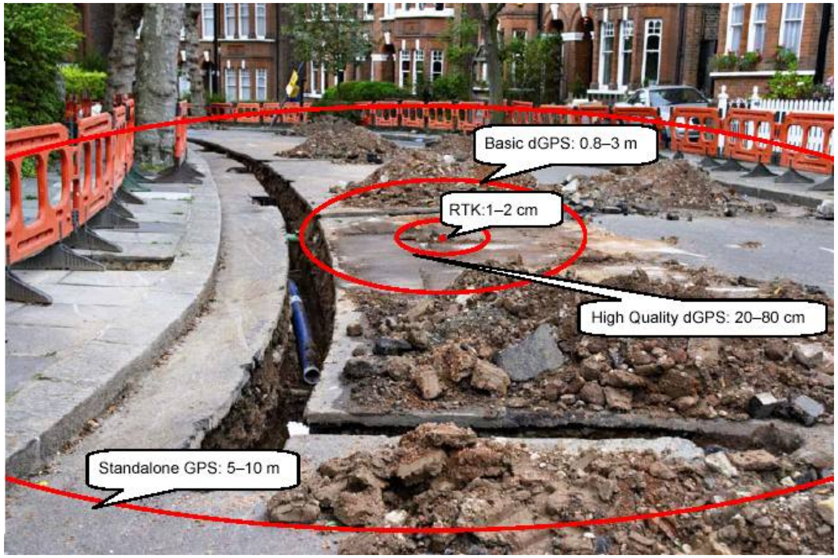
Objectives

Our goals were to develop:

- A simple statistical method for the control of non-normal errors.
- A suitable method for any error model (parametric or non-parametric).
- A method that runs on the population and not on parameters of the population.
- A method valid for 1D, 2D and 3D data and any kind of geometries (e.g. points, line strings, etc.).
- A method that allows to control the distribution of errors in several points (e.g. in the mean, in the mean + 1 deviation, etc.).

Contents

- Introduction
- The Normal distribution
- Non-normal error data (free-distributed)
- Multinomial approach
- Example
- Conclusion



Introduction

Importance

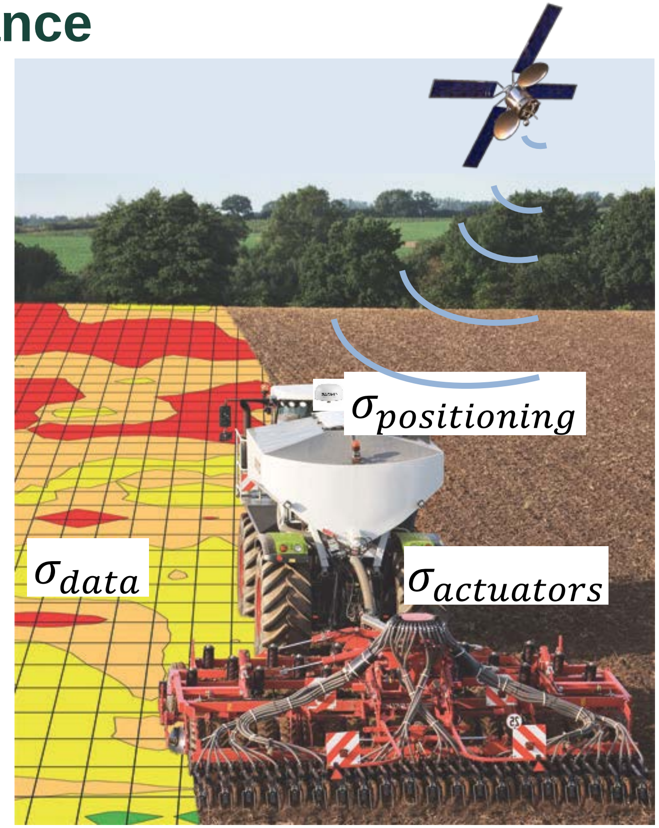
Positional accuracy is now of great importance

In general:

- Increase of use of GI implies increasing demand of quality.
- SDI need interoperability.
- GNSS allow everybody to get coordinates.

Demanding applications:

- Intelligence.
- Military applications (eg weapons and missiles)
- Unmanned vehicles (UA).
- Navigation.
- Precision farming.
- Etc.



Precision farming: seeder

Quality control is needed¡¡¡¡

Quality control method for non-normal positional error data using a multinomial approach

Introduction

PAAMs

There are many positional accuracy assessment methods (PAAMs) available:

- National Map Accuracy Standard (1947) by USBB.
- Accuracy Standards for Large Scale Maps (1990) by ASPRS
- Engineering Map Accuracy Standard (1983) by ASCE
- National Standard for Spatial Data Accuracy (1998) by FGDC
- STANAG 2215 by NATO.
- ASPRS Positional Accuracy Standards for Digital Geospatial Data (2014)
- Etc.
- UNE 148002.
- Etc.

norma española

UNE 148002

Septiembre 2014

TÍTULO | Información Geográfica

Control de calidad posicional de datos espaciales

United States National Map Accuracy Standards

With a view to the utmost economy and expedition in producing maps which fulfill not only the broad needs for standard or principal maps, but also the reasonable particular needs of individual

FGDC-STD-007.3-1998

NSDI

National Spatial Data Infrastructure

Geospatial Positioning Accuracy Standards
Part 3: National Standard for Spatial Data Accuracy

1068

ASPRS ACCURACY STANDARDS FOR LARGE-SCALE MAPS

The American Society for Photogrammetry and Remote Sensing
Approval by the ASPRS Professional Practicing Division, March, 1990

These standards have been developed by the Specifications and Standards Committee of the American Society for Photogrammetry and Remote Sensing (ASPRS). It is anticipated that these ASPRS standards may form the basis for revision of the U.S. National Map Accuracy Standards for both small-scale and

Table 1E — Planimetric Coordinate Accuracy Requirement (Ground X or Y in feet) for Well-defined Points - Class 1. Maps

PLANIMETRIC (X or Y) ACCURACY (limiting rms error, meters)	TYPICAL MAP SCALE

MINISTÈRE DE L'ÉQUIPEMENT, DES TRANSPORTS, DU LOGEMENT, DU TOURISME ET DE LA MER

Arrêté du 16 septembre 2003 portant sur les classes de précision applicables aux catégories de travaux topographiques réalisés par l'Etat, les collectivités locales et leurs établissements publics ou exécutés pour leur compte

NOR : *EQUP0300864A*

Le ministre de l'équipement, des transports, du logement, du tourisme et de la mer,

Vu le décret n° 92-706 du 21 juillet 1992 modifiant le décret n° 85-790 du 26 juillet 1985 relatif au rôle et à la composition du Conseil national de l'information géographique ;

Vu la loi n° 95-115 du 4 février 1995 d'orientation pour l'aménagement et le développement du territoire, modifiée par la loi

pour les deux coordonnées planimétriques et pour la coordonnée altimétrique). L'écart en position E_{pos} pour un point donné, par rapport à sa position issue d'un contrôle, est défini par la distance euclidienne, c'est-à-dire la racine carrée de la somme des carrés des écarts sur chacune des coordonnées soumise à la même classe de précision.

Une mesure n'est considérée comme mesure de contrôle que lorsque sont mis en œuvre des procédés fournissant une précision meilleure que celle de la classe de précision recherchée, avec un coefficient de sécurité C au moins égal à 2. C est le rapport entre la classe de précision des points à contrôler et celle des déterminations de contrôle, classe de précision qui est elle-même évaluée selon les règles de l'art. La taille et la composition de l'échantillon d'objets géographiques de contrôle sont précisées par contrat.

Introduction

PAAMs

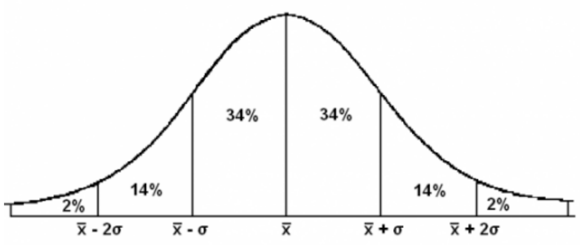
But PAAMs have problems:

Normality of errors (statistical model for the uncertainty)

A model is assumed in order to ease the analytical work and computations.

The assumed model is the Gaussian (NORMAL).

Some times explicitly and others implicitly.



Many studies point out that this hypothesis is not true

Other statistical models for the uncertainty:

- *LIDAR (Maune, 2007): non-parametric (distribution free)*
- *Manual digitizing (Bolstad et al 1990): Bimodal*
- *Digitizing (Tong & Liu, 2004): p-norm (Normal + Laplace)*
- *Geocodification (Cayo and Talbot 2003; Karimi and Durcik 2004, Whitsel et al. 2004): Log normal*
- *GNSS observations (Wilson, 2006; Logsdon, 1995): Raleigh, Weibull*
- *Other mentioned models are: Folder normal, Half normal, Gamma*

Introduction

PAAMs

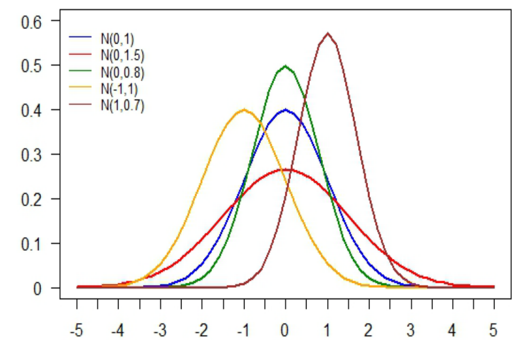
	NMAS	EMAS	ASLMS	NSSDA	STANAG	ISO 3951	ISO 2859
Issued by	UBB	ASCE	ASPRS	FGDC	NATO	ISO	ISO
Year	1947	1985	1990	1998	2002	2013	1999
Scale	All	>20000	>20000	All	<25000	"--"	"--"
RMSE based	No	No	Yes	Yes	No	No	No
Mean & Standard deviation based	No	Yes	No	No	Yes	Yes	No
Counting based	Yes	No	No	No	No	No	Yes
Implicit Normality of data	No	Yes	No	Yes	Yes	Yes	No
Control/**E**stimation	C	C	C	E	C	C	C
Isolated lot	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Lot by lot	No	No	No	No	No	Yes	Yes
Recommended sample size	--	>20	>20	>20	167	Variable	Variable
Known error type I	No	Yes	No	Yes	Yes	Yes	Yes
Known error type II	No	No	No	No	No	Yes	Yes

Normal distribution

The Normal distribution Is the basic distribution for error models.

- If each component of error follows a normal distribution model, all of them independents, we can assure that error is purely at random.
- We can see that a variable error, E_m is distributed according to a Normal distribution with parameters μ, σ, if its density function is:

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



- In this expression:
  - μ is the mean (of errors). If $\mu = 0$, there is no bias in the error distribution
  - σ is the standard deviation of errors. The greater the value of σ is, the more probable to find big errors is.
- The use of Normal distribution implies that errors have to have sign (positive-negative, left-right
- This model is required in the majority of PAAMs.

Normal distribution

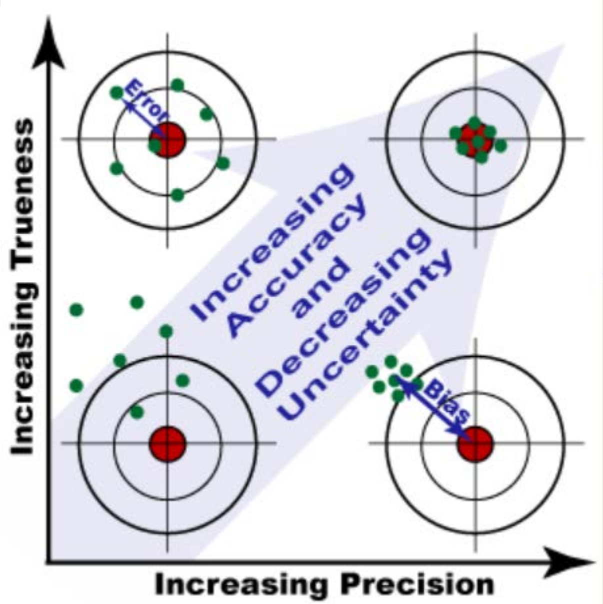
Accuracy: The closeness of agreement between a test result and the accepted reference value. [ISO 3534-1]

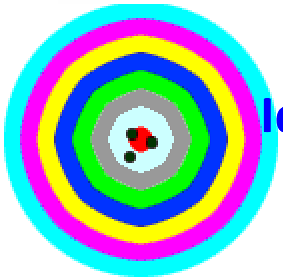
Accuracy = Trueness + precision

Bias component — Random component

Trueness: The closeness of agreement between the average value obtained from a large series of test results and an accepted reference value.

Precision: The closeness of agreement between independent test results obtained under stipulated conditions.





Ideal situation

The Normal Distribution is a parametric model well suited for accuracy:

- μ is related to bias in the error distribution.
- σ is the RMSE of the mean of the distribution.
- σ is the error for taking the mean as representation of the population.
- Mean = Mode = Median

Normal distribution

The parametric model is very convenient because it allows you to easily know the probabilities.
There is a direct relationship between certain expansion factors of the standard deviation and the probability. → Confidence intervals

$$IC(1-\alpha) = [\mu - K_{1-\alpha} \times \sigma/\sqrt{n}\,; \mu + K_{1-\alpha} \times \sigma/\sqrt{n}\,]$$

If $\mu=0$:

$$IC(1-\alpha) = [-K_{1-\alpha} \times \sigma/\sqrt{n}\,;\; K_{1-\alpha} \times \sigma/\sqrt{n}\,]$$

ISO 19157 measures

Table G.2 — Relation between the quantiles of the normal distribution and the significance level

Probability P	Quantile	Data quality basic measure	Name	Data quality value type
P = 50 %	$u_{50\%}$ 0,6745	$u_{50\%} \cdot \sigma_Z$	LE50	Measure
P = 68,3 %	$u_{68,3\%}$ = 1	$u_{68,3\%} \cdot \sigma_Z$	LE68.3	Measure
P = 90 %	$u_{90\%}$ = 1,645	$u_{90\%} \cdot \sigma_Z$	LE90	Measure
P = 95 %	$u_{95\%}$ = 1,960	$u_{95\%} \cdot \sigma_Z$	LE95	Measure
P = 99 %	$u_{99\%}$ = 2,576	$u_{99\%} \cdot \sigma_Z$	LE99	Measure
P = 99,8 %	$u_{99,8\%}$ = 3	$u_{99,8\%} \cdot \sigma_Z$	LE99.8	Measure

Result = MV $\pm K_\alpha \times \sigma_\mu$

Where:

MV → ***Mean value (usually 0)***

K_α → ***Quantile (e.g. 95%)***

σ_μ → ***Mean Standard deviation***

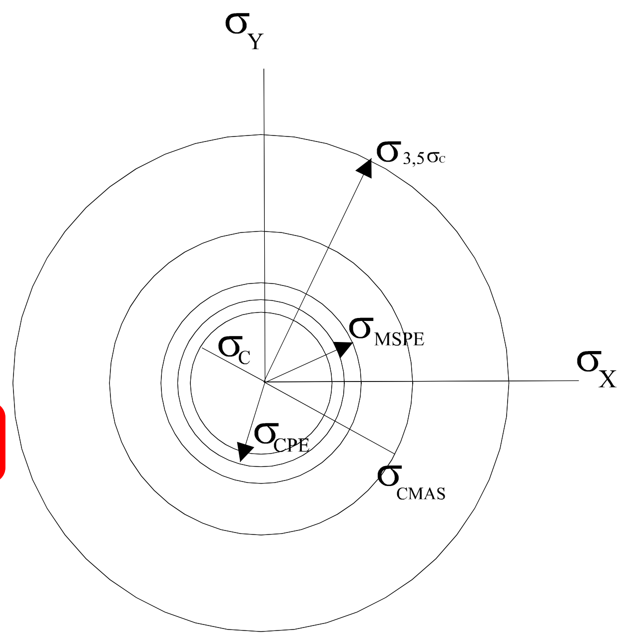
Quality control method for non-normal positional error data using a multinomial approach

Normal distribution

ISO 19157 measures

Table G.5 — Relationship between the probability P and the corresponding radius of the circular area

Probability *P*	Data quality basic measure	Name	Data quality value type
P = 39,4 %	$\frac{1}{\sqrt{2}}\sqrt{\sigma_x^2+\sigma_y^2}$	CE39.4	Measure
P = 50 %	$\frac{1,1774}{\sqrt{2}}\sqrt{\sigma_x^2+\sigma_y^2}$	CE50	Measure
P = 90 %	$\frac{2,146}{\sqrt{2}}\sqrt{\sigma_x^2+\sigma_y^2}$	CE90	Measure
P = 95 %	$\frac{2,4477}{\sqrt{2}}\sqrt{\sigma_x^2+\sigma_y^2}$	CE95	Measure
P = 99,8 %	$\frac{3,5}{\sqrt{2}}\sqrt{\sigma_x^2+\sigma_y^2}$	CE99.8	Measure



Name	Probability	Deviation
Circular standard error (σ_c)	0.3935	1.0 σ
Circular probable error (CPE, CEP)	0.5	1.1774 σ
Circular mean square positional error (MSPE)	0.6321	1.4142 σ
Circular map accuracy standard (CMAS)	0.9	2.1460 σ
Three-five sigma error (3.5σ)	0.9978	3.5 σ

Maling (1989)

Normal distribution

Why is important the normality of error data?

- The Normal distribution It is the distribution of pure random processes for continuous variables.
- The Normal distribution Is very adequate to be applied to the description of measurement errors
- The Normal distribution Is the underlying statistical model for the majority of statistical analysis.
- The Normal distribution Is the underlying hypothesis for the majority of the PAAMs.
- The model is easy to use and well-known.
- The model adequately models bias and dispersion, the two major concerns in spatial error analysis.

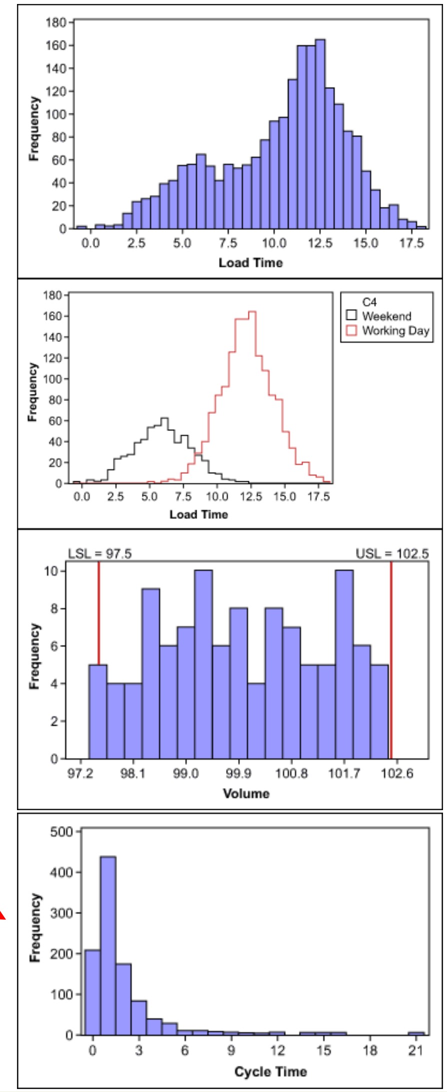
Normal distribution

Which is the origin of non-normal distributed data?

There are six main reason:

- Presence of too many extreme values
- Overlap of two or more processes
- Insufficient data discrimination (round-off errors, poor resolution)
- Elimination of data from the sample
- Values close to cero or natural limit
- Data follows a different distribution (e.g. Weibull, log-normal, exponential, Gamma, etc.)

One or more of these reasons can be present in our error data



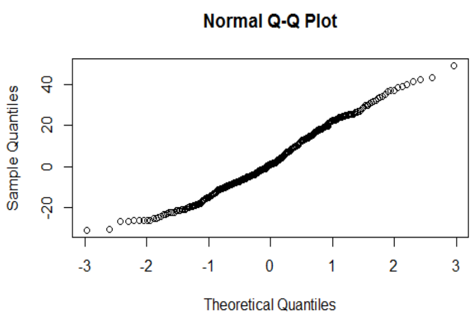
Normal distribution

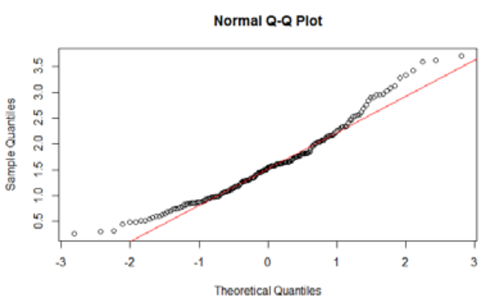
How can affect my assessment non-normal distributed data?

- If the underlying hypothesis is normality, non-normal distributed data means that **results are totally wrong**.

How can be verified?

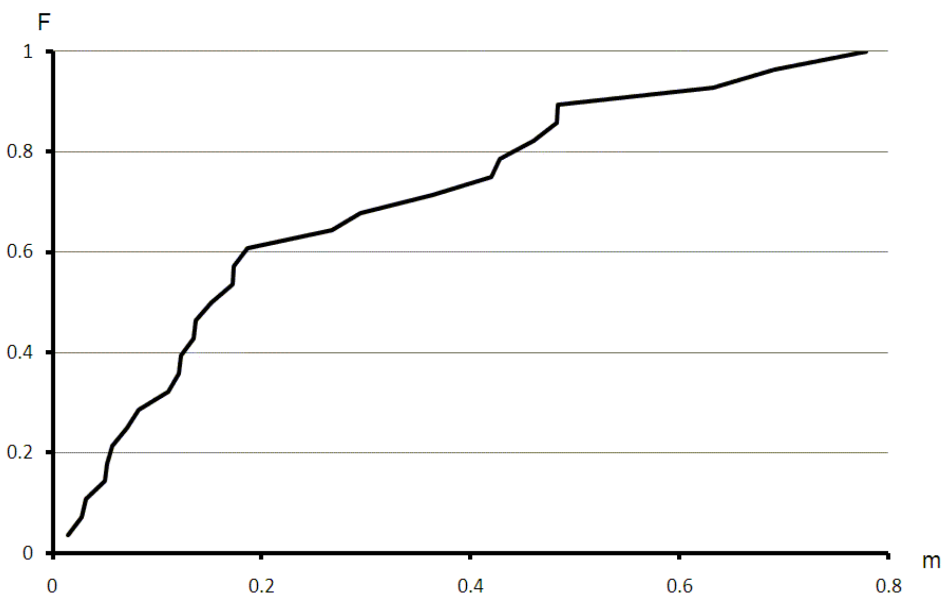
- They exist many statistical procedures to contrast this hypothesis.
- Some of them are: QQ-plot, Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors, Anderson–Darling, etc.

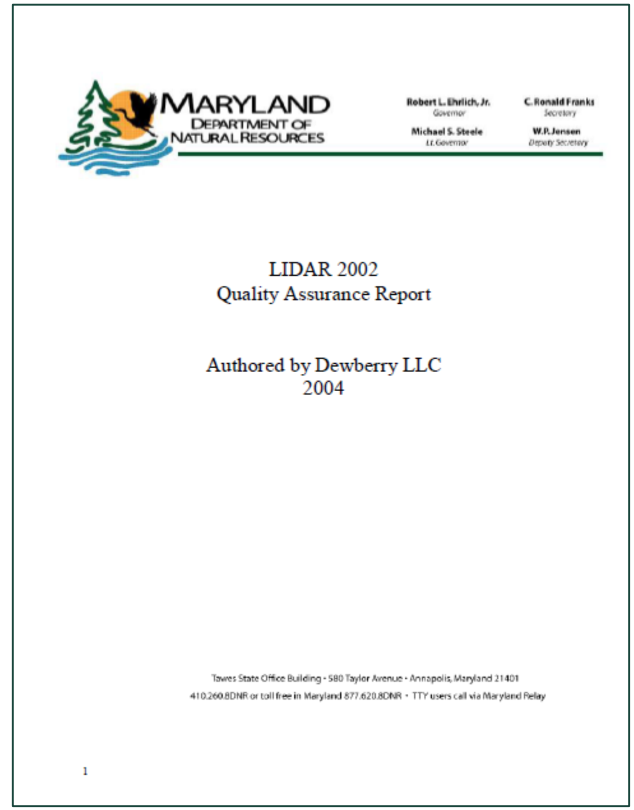




Non-normal error data (free-distributed)

Examples





MARYLAND
DEPARTMENT OF NATURAL RESOURCES

Robert L. Ehrlich, Jr.
Governor
Michael S. Steele
Lt. Governor
C. Ronald Franks
Secretary
W.P. Jensen
Deputy Secretary

LIDAR 2002
Quality Assurance Report

Authored by Dewberry LLC
2004

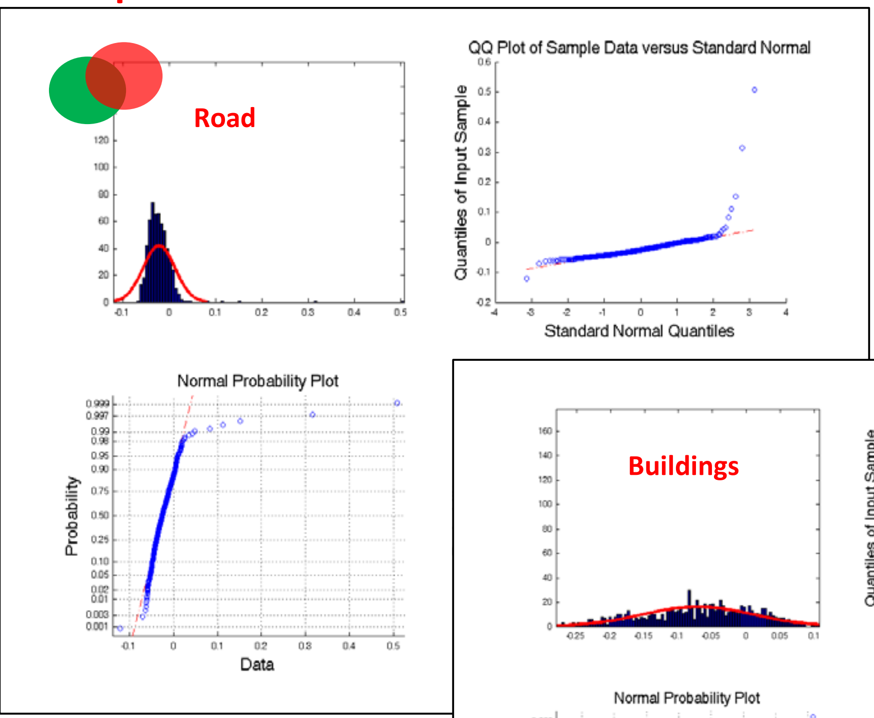
Tawes State Office Building • 580 Taylor Avenue • Annapolis, Maryland 21401
410.260.8DNR or toll free in Maryland 877.620.8DNR • TTY users call via Maryland Relay

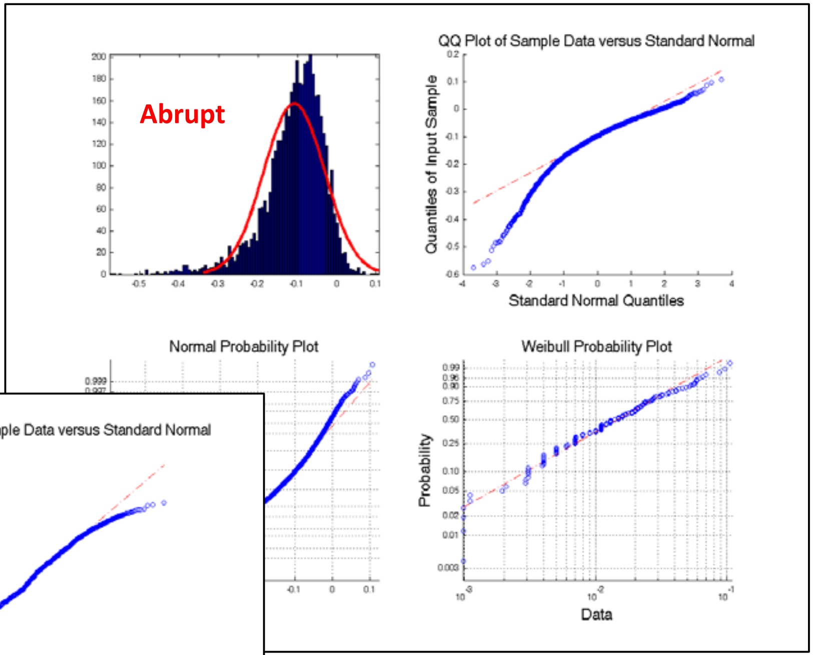
1

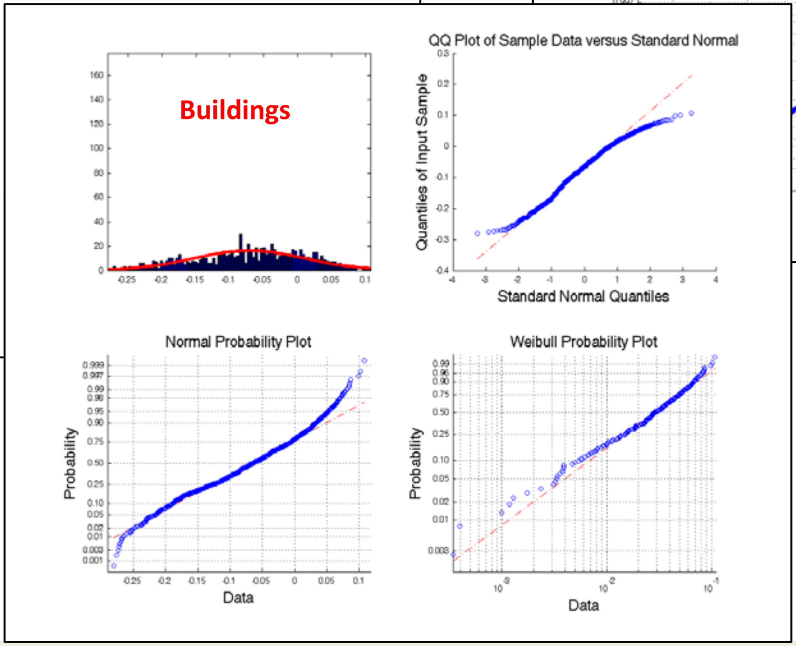
Source. Dewberry, (2004). Worcester County LIDAR 2002 Quality Assurance Report. Maryland Department of Natural Resources.

Non-normal error data (free-distributed)

Examples





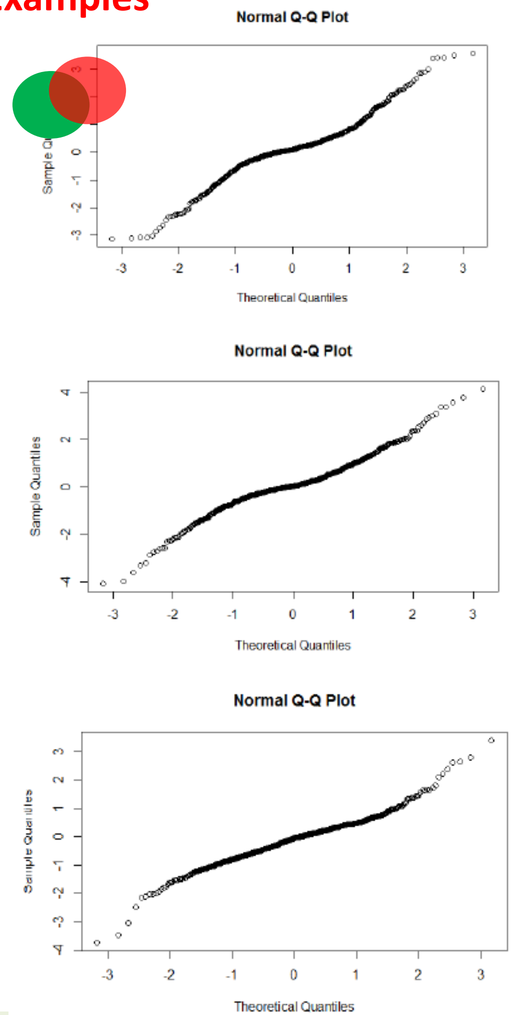


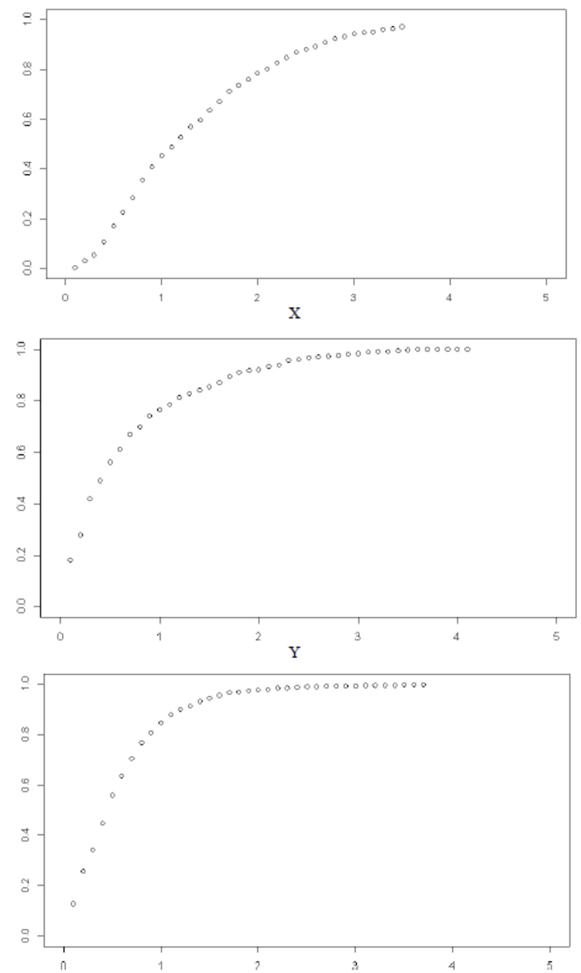
Avila

Quality control method for non-normal positional error data using a multinomial approach

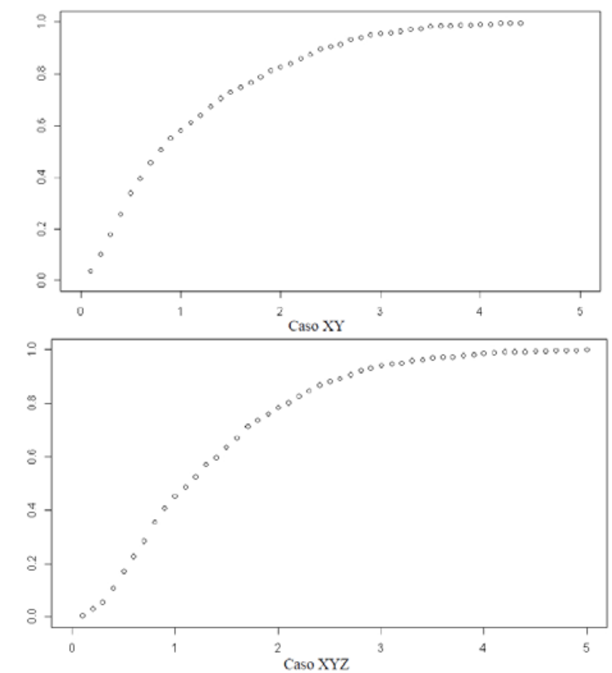
Non-normal error data (free-distributed)

Examples





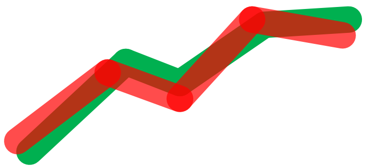
	Error en X		Error en Y		Error en Z
Mín.	:-3.1178	Mín.	:-4.06016	Mín.	:-3.72464
1st Qu.	:-0.2519	1st Qu.	:-0.33018	1st Qu.	:-0.57083
Median	: 0.0948	Median	: 0.04772	Median	:-0.04422
Mean	: 0.1135	Mean	: 0.10720	Mean	:-0.11497
3rd Qu.	: 0.5289	3rd Qu.	: 0.60150	3rd Qu.	: 0.32724
Max.	: 3.5674	Max.	: 4.14792	Max.	: 3.40110

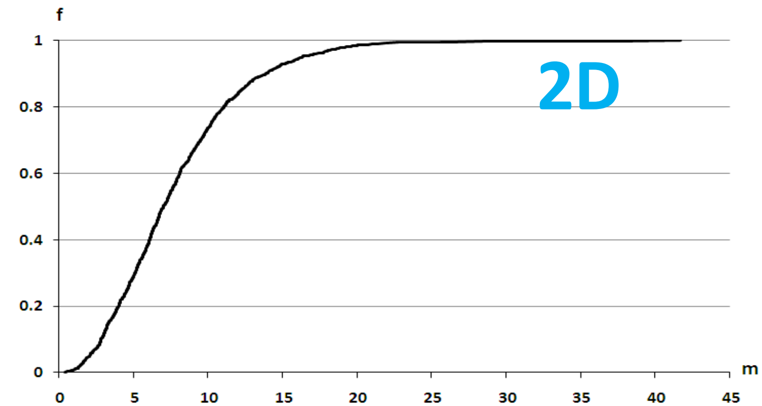


ICGC Institut Cartogràfic i Geològic de Catalunya

Non-normal error data (free-distributed)

Examples

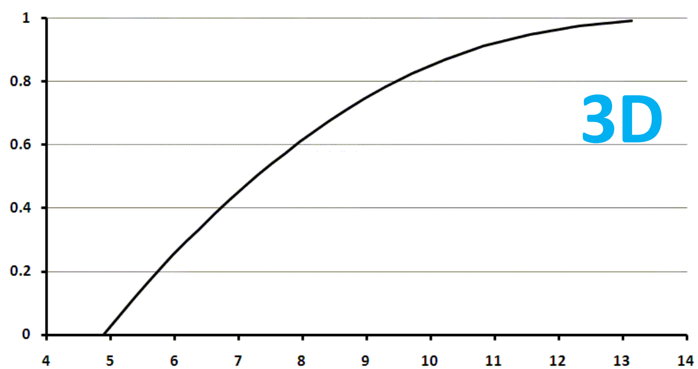




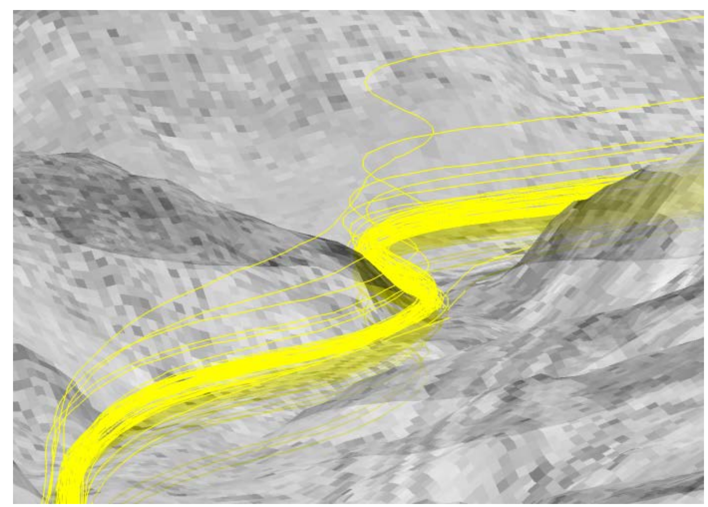
Principal characteristics of the product-axes and GPS-axes.

Characteristics	MTA10v product subset	GPS field survey
Total length	1210 Km	1210 Km
Total cases	1254 road segments	1254 road segments
Mean length	965 m	965 m
Standard deviation of the length	1671 m	1671 m
Total points involved	28,823 points	122,467 points
Mean points per road segment	22.98 points/road segment	97.66 points/road segment
Mean distance between points	41.98 m	9.88 m
Standard deviation of points distance	28.49 m	3.19 m
Mean speed kinematic survey	–	35.56 Km/h
Positional accuracy	10.65 m (95%)	1.41 m (95%)

(Hausdorff Distance)



Source: Ariza-López F.J, García-Balboa J.L, Ureña-Cámara M.A, Reinoso-Gordo F.J. (2012). Metodología para la evaluación de la calidad de elementos lineales 3D. En X Congreso TOPCART 2012, 16-19 Octubre, Madrid.



Non-normal error data (free-distributed)

Conclusion

Why are important methods for dealing with non-normal error data?

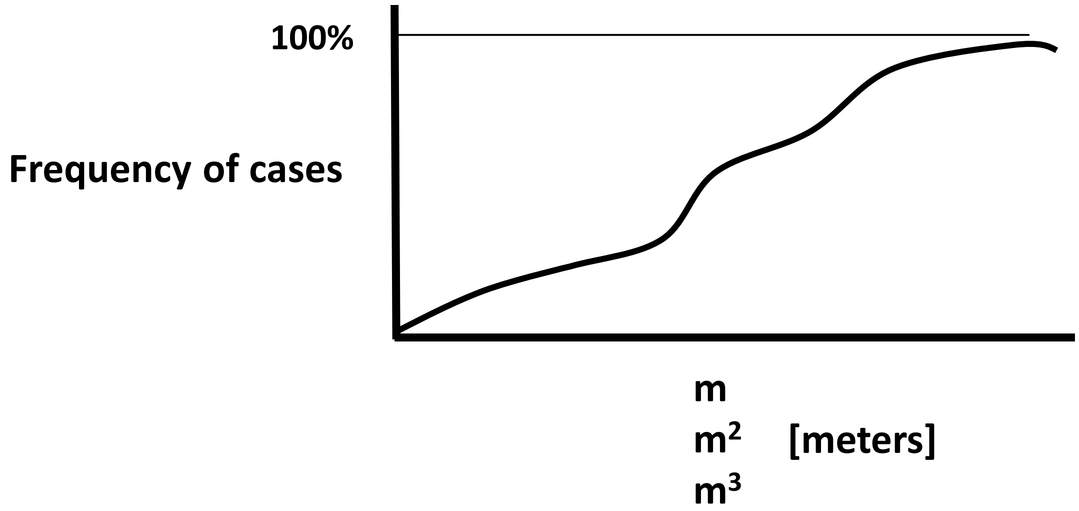
- It is a very common situation when dealing with spatial data (e.g. Lidar).
- For the majority of situations, non-normal error data means non-parametric models, so new methods are needed.
- In the Big-data parametric models are not so useful, it is possible to work with the observed model.

Multinomial approach

Positional defectives

Error: The difference between a measured value of a quantity and a reference value (conventional value or true value) [VIM 2007]

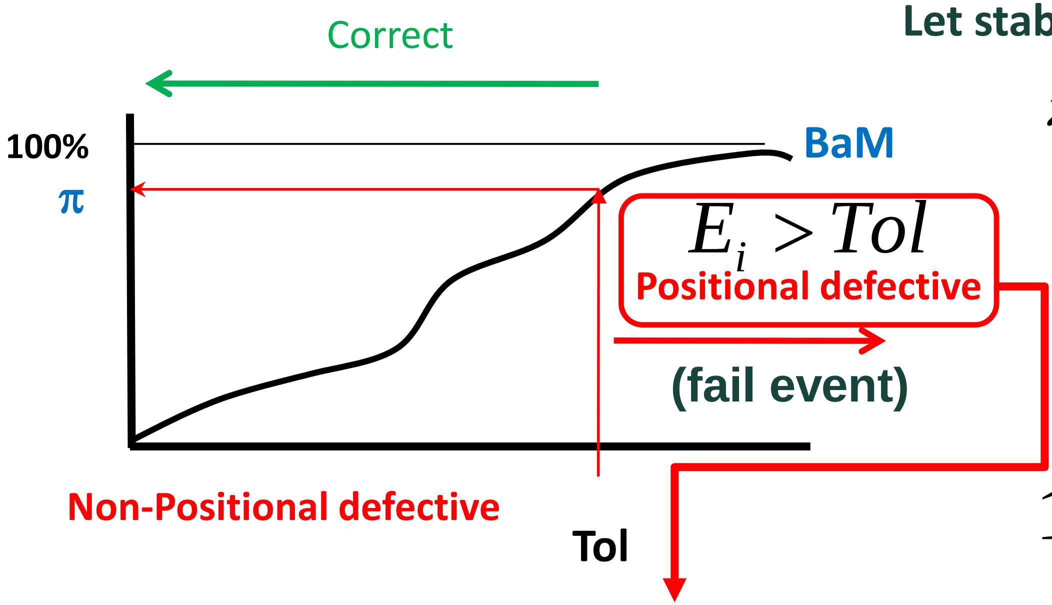
$$E_i = V_{Pro} - V_{Ref}$$



Observed distribution of errors

Base Model

Multinomial approach



Let stablish a metric tolerance Tol.

$$\pi = P[E_i > Tol]$$

π → proportion of positional defectives in the BaM established By the given Tol

$1-\pi$ → proportion of non-positional defectives in the BaM established By the given Tol

the cases to be counted¡¡¡

Binomial Distribution

$$P[F > mc \mid F \to B(n,\pi)] = \sum_{k=mc+1}^{n} \binom{n}{k} \pi^k (1-\pi)^{n-k}$$

$\mathrm{E}[X] = np$ $\qquad$ $\mathrm{V}[X] = npq$

Multinomial approach

Two tolerances

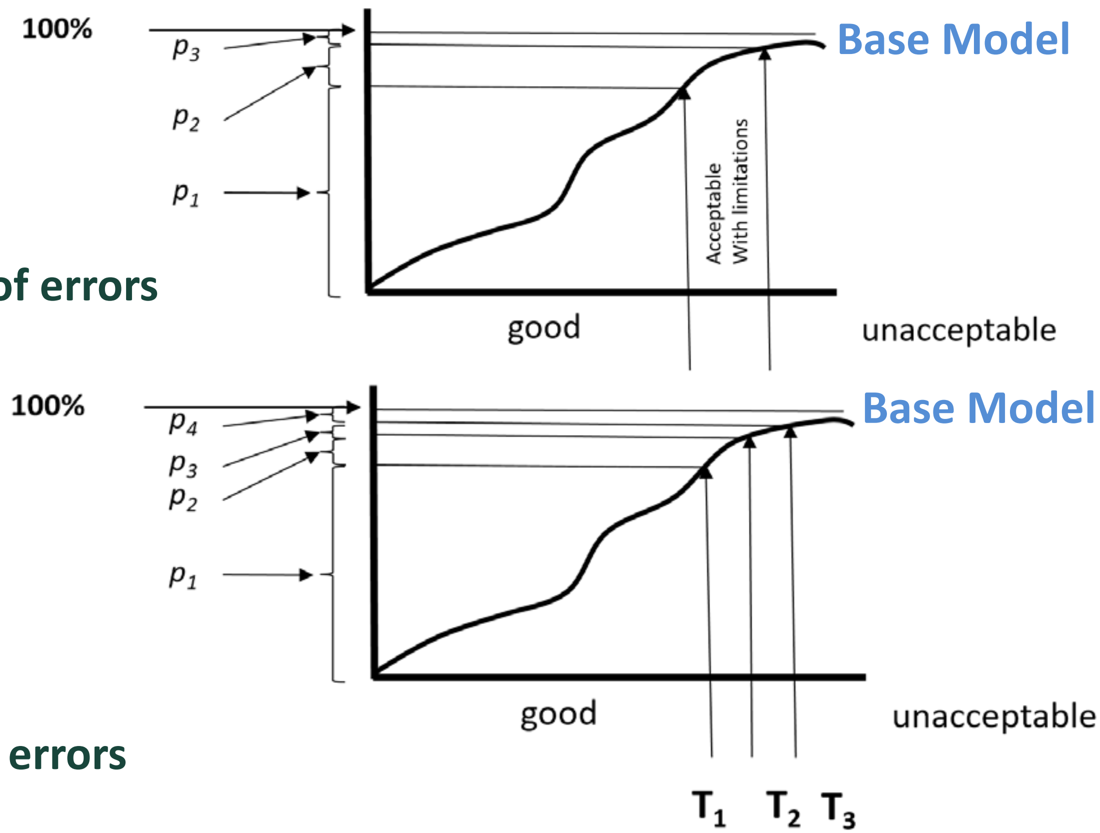
↓

Three categories of errors

Three tolerances

↓

Four categories of errors



k tolerances

Multinomial approach

Multinomial Distribution

- It is a multivariate extension of the Binomial Distribution
- It appears when the result of an experiment can be classified into $k > 1$ categories (When $k > 2$ we obtain the binomial distribution), and each of them with a probability π_i, $\pi_1 + \cdots + \pi_k = 1$.
- So, if an experiment is carried out N times, and the result is given by $(N_1, \ldots, N_k)$, the probability mass function is:

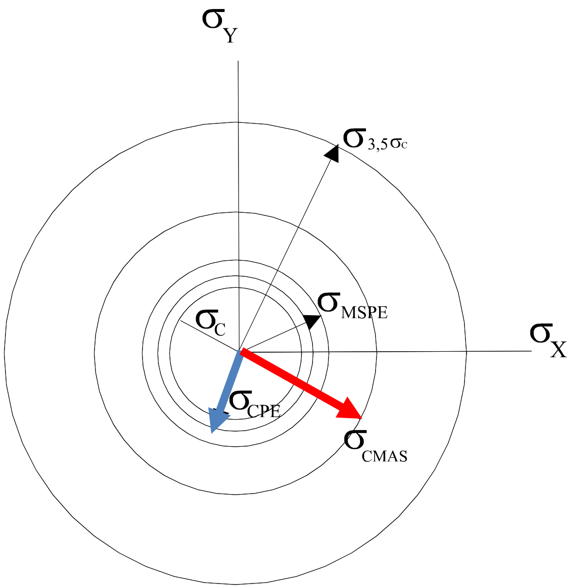
$$P[X_1 = N_1, \ldots, X_k = N_k] = \frac{N!}{N_1! \ldots N_k!} \pi_1^{N_1} \ldots \pi_k^{N_k}$$

Multinomial approach

The two tolerances case

Relation between the tolerances (T1 and T2) and the “Normal distribution case”

Name	Probability	Deviation
Circular standard error (σ_c)	0.3935	1.0 σ
Circular probable error (CPE, CEP)	0.5	1.1774 σ
Circular mean square positional error (MSPE)	0.6321	1.4142 σ
Circular map accuracy standard (CMAS)	0.9	2.1460 σ
Three-five sigma error (3.5σ)	0.9978	3.5 σ



Example:

Let be σ= 2m

Consider that we want to ensure that the distribution of observed errors meets, at least, the following two conditions:

- At least 50% of the errors is less than T1 (CPE).

- At least 90% of the errors is less than T2 (CMAS).

Product specifications

In this case a T_1 = 1.1774 x σ = 2,3548 m → Probability= 50%

In this case a T_2 = 2.1460 x σ = 4,2920 m →Probability = 90%

$$\pi_1 = 50\%$$
$$\pi_2 = 40\% \; (= 90\% - 50\%)$$
$$\pi_3 = 10\% \; (100\% - 90\%)$$

Multinomial approach

The two tolerances case

Let be two metric tolerances: T1 and T2

The following specifications (ISO 19131) for the product has been stated by the base model:

- The proportion of error cases where $E_i \leq T_1$ has to be equal or greater than π_1
- The proportion of error cases where $T_1 < E_i \leq T_2$ has to equal or be less than π_2
- The proportion of error cases $E_i > T_2$ has to be less than π_3

So we can classify the positional error E_i in a control element into three categories:

- small errors if $E_i \leq T_1$,
- moderate errors if $T_1 < E_i \leq T_2$, and
- excessive errors if $T_2 < E_i$.

To prove this a sample of size n is taken from a population of size N. So that:

- n_1 is the number of elements where Ei ≤ T1;
- n_2 is the number of elements with $T_1 < E_i \leq T_2$,
- n_3 the number of elements with $T_2 < E_i$.

Multinomial approach

The two tolerances case

In order to perform a hypothesis testing both a statistics and a null hypothesis are needed.

The statistic:
The sampling statistics is: $\boldsymbol{\nu}^* = (\boldsymbol{n_1}, \boldsymbol{n_2}, \boldsymbol{n_3})$, so that $\boldsymbol{n_1} + \boldsymbol{n_2} + \boldsymbol{n_3} = \mathbf{n}$.
The parameters of the multinomial distributions are: N, $\boldsymbol{\pi_1}, \boldsymbol{\pi_2}, \boldsymbol{\pi_3} = \mathbf{1} - \boldsymbol{\pi_1} - \boldsymbol{\pi_2}$.

The null hypothesis is:
- $\mathbb{H}_0$: The sampling statistics, $\boldsymbol{\nu}^*$, has a multinomial distribution with parameters $(\mathbf{n}, \boldsymbol{\pi}^0) = (\boldsymbol{\pi}_1^0, \boldsymbol{\pi}_2^0, \boldsymbol{\pi}_3^0)$ where $\boldsymbol{\pi}_k^0 = \boldsymbol{P}_k/100$ and $\boldsymbol{\pi}_1^0 + \boldsymbol{\pi}_2^0 + \boldsymbol{\pi}_3^0 = \mathbf{1}$.

− $\mathbb{H}_1$: The alternative hypothesis is that the true distribution of errors presents more large errors than the specified under $\mathbb{H}_0$ → At least one of these conditions: $\pi_1 \geq \pi_1^0$ or $\pi_2 \leq \pi_2^0$, or $\pi_3 \leq \pi_3^0$, is false. Here the alternative hypothesis specifies what we consider a worse situation, and this situation takes place when the proportion of elements with tolerance less than T_1 is less than P_1, or when the other two proportions account for more than P_2 or P_3, because this implies a worsening in tails.

Multinomial approach

The two tolerances case

P-value: This is an exact test, so the p–value is calculated as follows:
Given the test statistics $\nu^* = (n_1, n_2, n_3)$ we calculate the probability in the multinomial fixed by the null hypothesis to the obtained value and those counting of elements $\mathrm{m} = (m_1,\ \ m_2,\ m_3)$ that verify:

- $m_1 < n_1$
- $m_1 = m_1$ and $m_2 \leq n_2$

Adding up the p-values of all the cases that verify these conditions and rejecting the null hypothesis if the p-value obtained (the sum) is less than α

Example

A proof

Let be a product specification where T1= 9.243 m^2, T2= 14.065 m^2

And Positional errors in X, Y and Z are considered to be distributed according to three Normal and independent distributions with μ=0 m and σ=1.5 m (Base Model).→ This is our $\mathbb{H}_0$.

In this case, the composed quadratic error:

$$QE_i = Ex_i^2 + Ey_i^2 + Ez_i^2$$

is distributed according to a Gamma distribution with parameters of shape K=3/2 and scale θ=4.5.

For this parametric model we know that:

- The probability that an element has a $QE \leq 9.243$ m^2 is 0.75
- The probability that an element has a $QE \leq 14.065$ m^2 is 0.90.

In consequence, the error-cases quantities will follow the multinomial: *M*(n, 0.75, 0.15, 0.10).

Now, let consider the following three cases:

- C#1. $\mathbb{H}_0$ is true.
- C#2. The true model of the data errors is worse → there are higher number of positional defectives in p2 and p3.
- C#3. The true model of the data errors is better → there are less number of positional defectives in p2 and p3.

Quality control method for non-normal positional error data using a multinomial approach

Example

A proof

The symbol "▫" means $QE \leq 9.243\ \mathrm{m}^2$, the symbol "▪" means $9.243\ m^2 \leq QE \leq 14.065\ m^2$, and the symbol "■" means $QE > 14.065\ \mathrm{m}^2$

$C\#1 \rightarrow v^* = (15, 4, 1)$

C#1 Case of the hypothesis N(μ=0, σ=1,5)				C#2 Worse than the hypothesis N(μ=0, σ=2)				C#3 better than the hypothesis N(μ=0, σ=1)			
e_X[m]	e_Y[m]	e_Z[m]	T	e_X[m]	e_Y[m]	e_Z[m]	T	e_X[m]	e_Y[m]	e_Z[m]	T
-0,371	-1,672	2,755	▪	4,263	2,439	3,298	■	0,745	-0,001	-0,892	▫
-3,359	-0,815	1,454	▪	-2,547	-0,959	-0,483	▫	-1,174	-0,299	0,527	▫
0,251	0,340	0,467	▫	0,876	3,985	-0,851	■	0,938	0,031	0,993	▫
-0,308	-0,324	0,718	▫	-0,010	0,352	2,098	▫	0,219	-1,092	0,651	▫
1,172	-0,320	-0,411	▫	-1,920	0,744	4,166	■	-1,533	-2,152	-1,834	▪
-0,206	-3,074	-1,651	▪	2,313	-1,372	-3,335	■	0,481	-0,010	0,497	▫
3,873	0,304	2,280	■	-2,584	2,832	0,049	■	-1,551	-0,163	0,902	▫
0,394	-0,442	0,989	▫	0,830	-0,932	-0,510	▫	-0,383	0,239	-1,118	▫
2,322	-1,667	0,623	▫	1,895	-0,902	-2,230	▪	-1,267	2,032	-0,887	▫
-1,380	-2,260	0,342	▫	2,242	-1,206	1,741	▪	1,555	2,436	-0,998	▪
1,384	-1,444	-1,730	▫	-1,341	1,723	-0,745	▫	-0,371	-0,219	1,323	▫
-1,131	-0,549	-0,930	▫	-1,457	-1,699	-4,995	■	-0,217	0,438	0,003	▫
0,423	0,627	-1,257	▫	-0,541	4,164	1,924	■	1,606	-1,278	-0,310	▫
1,494	-1,359	-2,168	▫	2,818	2,699	-0,834	■	-1,338	-0,733	0,132	▫
-1,740	0,017	-1,281	▫	0,772	-0,099	-2,907	▫	-0,365	1,711	0,526	▫
-1,397	-0,196	0,214	▫	3,217	-1,191	1,744	■	-1,115	-1,208	-0,971	▫
1,670	0,262	-2,015	▫	0,343	-0,024	4,905	■	0,004	-0,203	0,307	▫
-0,399	-1,194	1,553	▫	-4,844	0,044	0,493	■	-1,031	0,998	-0,232	▫
-0,309	-1,106	1,562	▫	-0,050	-0,657	0,206	▫	0,740	-0,638	-0,397	▫
-1,329	0,014	2,745	▪	-1,096	-1,909	-1,731	▫	0,861	-0,080	0,879	▫

Quality control method for non-normal positional error data using a multinomial approach

Example

A proof

P-value for the exact test:

$C\#1 \rightarrow \boldsymbol{\nu}^* = (15, 4, 1)$

dmultinom (c(15,4,1), size=20, c(0.75,0.15,0.10))

dmultinom (c(13,6,1), size=20, c(0.75,0.15,0.10))

dmultinom (c(0,0,20), size=20, c(0.75,0.15,0.10))

C#1

$m_1+m_2+m_3=20$

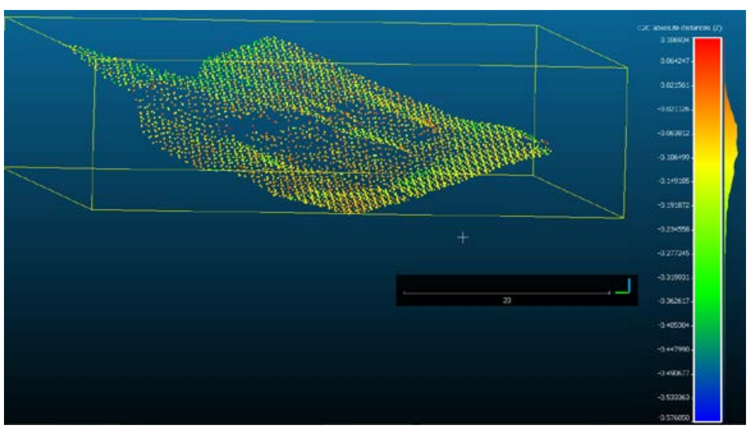
Worse value	m_1	m_2	m_3	Probability	Accumulated probability
1	15	4	1	0,05244	0,05244
2	15	3	2	0,06992	0,12236
3	15	2	3	0,04661	0,16898
4	15	1	4	0,01553	0,18452
5	15	0	5	0,00207	0,18659
6	14	6	0	0,00786	0,19446
7	14	5	1	0,03146	0,22593
8	14	4	2	0,05244	0,27837
9	14	3	3	0,04661	0,32499
10	14	2	4	0,02330	0,34830
11	14	1	5	0,00621	0,35451
12	14	0	6	0,00069	0,35520
13	13	7	0	0,00314	0,35835
14	13	6	1	0,01468	0,37303
….	….	….	….	….	….
….	….	….	….	….	….
….	….	….	….	….	….
195	0	20	0	3,32E-17	0,56942
196	0	19	1	4,43E-16	0,56942
….	….	….	….	….	….
214	0	1	19	3E-19	0,56942
215	0	0	20	1E-20	0,56942 = p-value

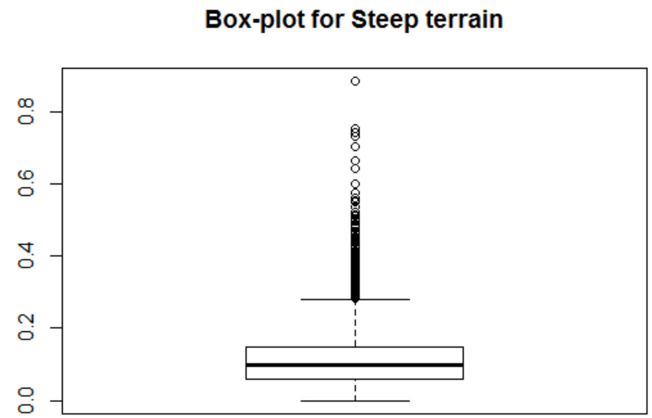
Example

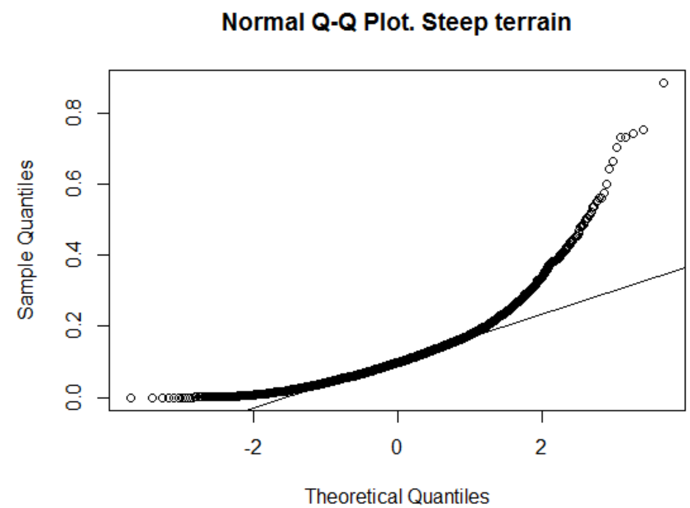
Lidar data **Base Model**

	Steep terrain
Minimum	0.00000
1st Quartile	0.05798
Median	0.09802
Mean	0.11341
3rd Quartile	0.14697
Maximum	0.88501

N = 190000, n = 4500







Example

Lidar data

Considering this Base Model, we are going to proof the proposal for two cases:

- Case A: Null hypothesis is true
- Case B: Null hypothesis is false

Method → Simulation procedure:

- 2000 samples of sizes 20 and 60 are taken
- In each sample:
  - i. The estimator ν^* is calculated, counting the number of points whose value falls in each category
  - ii. The p-value is calculated applying the procedure above described (slide #).

Example

Lidar data

Case A: Null hypothesis is true

I. At least the 80% of points present a value less than 0.161

II. Only the 5% of points present a value greater than 0,264

The proportion of times where the null hypothesis is rejected has to be approximately equal to the value of α proposed (Null hypothesis true)

Alpha value	% of rejected samples	
	N=20	N=60
10%	9.45%	9.76%
5%	4.84%	4.35%
1%	1.12%	0.9%

Example

Lidar data

Case B: Null hypothesis is false

I. At least the 80% of points present a value less than 0.15
II. Only the 5% of points present a value greater than 0.25

The proportion of times where the null hypothesis is rejected has to be greater than the value of α proposed (Null hypothesis false)

Alpha value	% of rejected samples	
	N=20	N=60
10%	19.75%	29.70%
5%	11.05%	18.35%
1%	3.94%	5.25%

Conclusions

- A new statistical method for positional control has been presented.
- The method is simple and has a well-founded statistical base.
- This method can be applied to any kind of error model (parametric or non-parametric) and to any kind of geometry (e.g. points, line strings, etc.)
- This method can be applied to cases of any dimension (1D, 2D, 3D, …nD)
- The method allows to control the distribution of errors in several points.
- The main strengths are:

  -It is not linked to any specific statistical hypothesis on errors.

  -Flexible in order to stablish the metric tolerances.

  -Metric tolerances can be related to standard parametric error models.

EuroGeographics
EuroSDR
ICA ACI
OGC Open Geospatial Consortium, Inc.
ISO

Malta, February 2018

2nd INTERNATIONAL WORKSHOP ON SPATIAL DATA QUALITY

Quality control method for non-normal positional error data using a multinomial approach

Thank you

Francisco Javier Ariza-López, José Rodríguez-Avi ,Virtudes Alba-Fernández, José Luis García-Balboa

[fjariza, jravi, mvalba, jlbalboa]@ujaen.es; chamorro@us.es

Universidad de Jaén; Paraje de las Lagunillas S/N, E-23.071-Jaén

UNIVERSITAS GIENNENSIS

Funded by the Project: MEMCALIG (Modelización Estadística de Matrices de Confusión en Calidad de la Información Geográfica)

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

PROYECTO COFINANCIADO POR LA UNIÓN EUROPEA
FONDO EUROPEO DE DESARROLLO REGIONAL