Simulation for estimating the probability of positional errors greater than a tolerance in a mixture of parametric models

Ariza-López, FJ[1]*; Rodríguez-Avi, J*[1]

[1] Universidad de Jaén (España).

Abstract

The control of the positional quality of a spatial product can be conducted by counting number of points of a population that will have an error greater than a given tolerance, which may be imposed as a null hypothesis for any model. This paper presents a method of using statistical simulation in order to apply this proposal to any mixture of parametric base models. The conceptual base, the application procedure and examples are presented.

Keywords: positional accuracy, binomial distribution, tolerance.

1. Introduction

One of first positional accuracy evaluation procedures is the National Map Accuracy Standards (NMAS), which proposes obtaining a random sample of size *n* and counting the number of sampling points for which its error (1) is greater than the specified tolerance. If this number is greater than 10% of *n*, the product will be rejected.

NMAS is an easy procedure, but we cannot employ it as a statistical test of hypothesis because we cannot control the type I error. But we can see this procedure as follows (Ariza-López and Rodríguez-Avi, 2014a): the probability of rejecting the null hypothesis –the product is adequate– is related to a Binomial distribution with parameters n –the sample size– and π –the population probability that a point has an error greater than the tolerance–. In consequence, if π were known, we could calculate the probability of rejecting the null hypothesis.

We define the error of a point as:

$$E_i = \sqrt{\sum_{j=1}^{p}\left(x_{ij} - x_{ij}^{R}\right)^2} \qquad (1)$$

where p is the number of dimensions (e.g. p=2 for 2D data) and $x_i = (x_{ij}); x_i^R = (x_{ij}^R),$ are respectively the product and true position of a control point.

The second step is to employ a Binomial distribution with parameters n –the sample size- and π. For this reason we have to determine the population distribution of errors that we want to contrast (the base model), that is to say

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

In Ariza-López and Rodríguez-Avi (2014b) this proposal is applied to non parametric models. In this paper we present a method of using statistical simulation in order to apply this proposal to any mixture of parametric base models. In this case the proposal is of interest when errors are formulated as complex mixtures of parametric models (e.g. different standard deviations in each positional component, correlations, etc.). The simulation method is also of interest to non-parametric cases, because its application can be extended easily to observed distributions.

2. Distribution of E_i

The problem is to determinate the true distribution function of E_i . We have to take into account that E_i is not directly observable, and, even if we could estimate parameters of the model by a likelihood method, the explicit form of the probabilistic distribution of E_i would be impossible to know.

Only under specific circumstances in distributions assumed for each error coordinate (e.g. X, Y, Z) in the null hypothesis we can obtain the exact value of π. These cases are:

- If errors in all coordinates are independent and normally distributed with μ=0 and $\sigma^2 = 1$, then E_i^2 has a Chi-square distribution with 2 degrees of freedom for the 2D case and 3 d.f. for the 3D case
- If errors are independent and normally distributed with μ=0 and equal variance σ^2, then E_i^2 has a Gamma distribution with parameters 1 (2D case) or 3/2 (3D case) and $2\sigma^2$

For any other case (e.g. non-equal variances, presence of correlated errors, no-normality, etc.) the exact probability distribution is not easy to calculate and, in consequence, our proposal is to apply a simulation procedure.

3. Approximate distribution of E_i

We can obtain π for any parametric case in the distribution of errors for each component through an approximation procedure. We are building a test of hypothesis, so the distribution of errors under the null hypothesis has to be completely known. This distribution of errors is, for instance, the distribution that we consider as adequate, and where we can establish correlated errors, different means and/or variances.

Once we have fixed this theoretical distribution we proceed as follows:

1. With a simulation procedure, we built m populations of size s. For instance, we propose m=5000 and n = 20000.
2. For each simulated point we calculate the quadratic mean error (QME), and for each population, j, we estimate the value of π_j as the number of points with a QME greater than the fixed tolerance divided by s.
3. The population value of π is the sum of π_j divided by m.

Through this algorithm we can obtain an adequate value of π, and it may be easily implemented, for instance, in the statistical software R. We want to remark that this procedure is valid, even when spatial autocorrelation exists in the point distribution

4. Application examples.

We present several illustrative examples. Firstly we compare the estimated value of π with the exact one when we know the distribution (cases "a" and "b"), both cases are presented in order to show that the simulation procedure obtain similar values that the theoretic distributions. After we propose a mixture of models (case "c").

a) Case 2D with errors in X →N(0,1), Y →N(0, 1) independent errors. We propose the use of the R function *rmvnorm* to generate normal multivariate populations. The exact distribution is χ^2 with 2 d.f. Table 1 shows results obtained through the algorithm and the exact results.

Table 1: Probability of quadratic error greater than a tolerance (Chi2 model).

Tolerance (m)	Value of π (estimated)	χ^2_2
1	0.6064349	0.6065306
2	0.1353595	0.1353352
3	0.0111034	0.0111089
4	0.0003308	0.0003354

b) Case 3D with errors in X →N(0,2.5), Y →N(0, 2.5), Z →N(0, 2.5) independent errors. The exact distribution is Gamma (3/2, 12.5). Results are shown in Table 2.

Table 2: Probability of quadratic error greater than a tolerance (Gamma(3/2, 12.5) model).

Tolerance (m)	Value of π (estimated)	Gamma(3/2, 12.5)	Tolerance (m)	Value of π (estimated)	Gamma(3/2, 12.5)
1	0.983782	0.983773	5	0.261448	0.261464
2	0.887233	0.887217	6	0.123937	0.123889
3	0.696179	0.696186	7	0.049428	0.049437
4	0.464531	0.464545	8	0.016622	0.016632

c) 3D case with errors in X →N(0,1), Y →N(0, 2) but with a planimetric correlation ρ_{xy}=0.5, and Z → N(0,4) independent of X and Y. In this case, we cannot know the exact distribution, but simulation allows us to estimate the probabilities (Table 3):

Table 3: Probability of quadratic error greater than a tolerance (mixture of three normal distributions).

Tolerance (m)	Value of π (estimated)	Tolerance (m)	Value of π (estimated)
1	0.967341	5	0.281756
2	0.823473	6	0.173583
3	0.623635	7	0.101999
4	0.433222	8	0.057059

Now consider a positional quality control of a spatial dataset whose positional errors follow the previous case "c". In addition to the corresponding sample of size n, we can use the adequate value of π as the second parameter in a Binomial in order to make the pass/fail decision. So, under the hypothesis of case "c"if in a sample of size n=20 we obtain F=5 points whose $E_i > 7m$, the corresponding p-value is given by:

$$P[F \geq 5] = 1 - \sum_{j=0}^{4} \binom{20}{j} 0.101999^{j} (1-0.101999)^{20-j} = 0.0464$$

That is to say, $P[F \geq 5]$ =0.0464 in a B(20, 0.101999), and we reject the hypothesis that this product verifies conditions given in case "c" when significance is 5% (α=0.05).

5. Conclusion

A procedure for obtaining population probabilities for the QME for positional control has been presented. As a result the π parameter for the Binomial distribution allows us to obtain the p-value to contrast the null hypothesis that the product has the proposed error distribution.

Examples demonstrate the general validity of the procedure for parametric models of errors by comparing the exact with the approximate values, and the ability of the procedure for mixtures of parametric base models, even when spatial autocorrelation appears.

Acknowledgments

This work has been funded by the Ministry of Science and Technology (Spain) and the European Regional Development Fund under grant no. BIA2011-23217. The authors also acknowledge the Regional Government of Andalusia (Spain) for the financial support since 1997 for the research groups with code PAIDI-TEP-164 and PAIDI-FQM-245.

References

Ariza-López FJ, Rodríguez-Avi J (2014a) A Statistical Model Inspired by the National Map Accuracy Standard. Photogrammetric Engineering & Remote Sensing Vol. 80 (3), 271-281.

Ariza-López FJ, Rodríguez-Avi J (2014b) A Statistical Model for Positional Quality Control of Spatial Data. Spatial Accuracy 2014, East Lansing, Michigan, July 8-11.

R Development Core Team. (2009). R: A language and environment for statistical computing. -Vienna, Austria: R Foundation for Statistical Computing.