A Meta-classifier Approach for Outlier Identification in Geodetic Networks

Stefano S. Suraci[1], Ronaldo R. Goldschmidt[1], Leonardo C. Oliveira[1], Ivandro Klein[2]

[1]Defense Engineering Program – Instituto Militar de Engenharia (IME)
22.290-270 – Rio de Janeiro – RJ – Brazil

[2]Graduate Program in Geodetic Sciences – Universidade Federal do Paraná (UFPR)
81.531-990 – Curitiba – PR - Brazil

{stefano,ronaldo.rgolg,leonardo}@ime.eb.br, ivandro.klein@ifsc.edu.br

Abstract. *Geodetic networks provide the positional basis for geospatial data collection. Outlier identification is a routinely task in quality control of geodetic networks. In this work, we proposed a meta-classifier approach for outlier identification in a leveling network. Based on Monte Carlo methods, we created a database combining results and features originally from both Iterative Data Snooping and Minimum L1-norm, usual methods of outlier identification in geodetic networks. Promising results pointed a higher accuracy of our Multilayer Perceptron-based meta-classifier in relation to them. In addition to many relevant points raised for future work, these results highlight the potential of this research and justify its continuity.*

1. Introduction

A geodetic network is a set of points with high precision coordinates that materialize a reference system. Geodetic networks provide the positional basis for geospatial data collection. Production and quality assessment of geospatial datasets highly relies on data and coordinates derived from geodetic networks. If the reference geodetic network coordinates (horizontal and/or vertical) are not of appropriate quality, there is a potential for high positional error propagation in the geoinformation derived from it.

Being m and n the quantity of observations and parameters (unknowns) respectively, $\mathbf{A_{mxn}}$ the design matrix, $\mathbf{x_{nx1}}$ the vector of unknowns, $\mathbf{l_{mx1}}$ the vector of observed values, $\mathbf{v_{mx1}}$ the observational residuals vector, $\mathbf{\Sigma_l}$ the covariance matrix of observations, σ_0^2 the *a-priori* variance factor of unit weight and $\mathbf{P_{mxm}}$ the weight matrix of observations, the mathematical model of a geodetic network may be defined as:

$$Ax = l + v;\; P = \sigma_0^2 * \Sigma_l^{-1} \quad (1)$$

The Least Squares (LS) is the default method for the adjustment computation of geodetic observations. It minimizes the sum of the weighted squared residuals:

$$LS\colon v^T P v = min \quad (2)$$

However, LS parameters estimation is of poor quality when outliers are present in the set of observations (Ghilani 2010). Hence, outlier identification is a routinely task in quality control of geodetic networks (Klein et al. 2021). Two main approaches arise for outlier identification in geodetic networks: tests based on LS residuals and other robust estimation criteria. About the first approach, as a variation of the pioneering Data

Snoping procedure of Baarda (1968), the Iterative Data Snooping (IDS) (Teunissen 2006) is the best-established outlier identification procedure in the geodetic literature.

In IDS, considering LS results and independent geodetic observations, if the maximum absolute normalized residual is higher than a user-controlled critical value w_{IDS}, the respective observation is classified as outlier and usually removed from the sample of observations. This procedure is repeated iteratively until no outlier is found. Being v_i the observation residual and σ_{v_i} its respective standard-deviation, the absolute normalized residual of i^{th} observation is given by:

$$w_i = \frac{|v_i|}{\sigma_{v_i}} \tag{3}$$

Differently from LS, robust estimators are in general more "resistant" to outliers. Minimum L1-norm (ML1) is one of the standard robust estimation methods in geodetic networks. Being **p** the weights vector of independent observations, the ML1 deals with the minimization of the sum of weighted absolute residuals:

$$ML1: p^T|v| = min \tag{4}$$

Alternatively, the identification of outliers under ML1 tends to provide higher accuracy if all observations are assumed to have the same weight (Suraci et al. 2019). In this article, we are calling the "regular" case as Weighted Minimum L1-norm (WML1), and the latter simplification of the weights as Simplified Minimum L1-norm (SML1). Note that Equation 4 is also valid for SML1 by making all elements of **p** equal to "1".

In any case, ML1 may consider different criteria for outlier identification. Here we follow that of Amiri-Simkooei (2018): all (if any) observations with absolute normalized residuals higher than a user controlled critical value w_{ML1} are classified as outliers. Note that differently from IDS, Minimum L1-norm is not iterative, i.e., outliers are identified simultaneously instead of one at a time. Here, we will call the critical value w_{WML1} when performing WML1 and w_{SML1} when performing SML1.

Meta-classifiers are classification models that aim to integrate multiple independently obtained base classifiers (Goldschmidt et al. 2015). As is desired in meta-classification, IDS and WML1 accuracies in outlier identification are diverse (to some extend), since they vary differently with network redundancy: $r=m-n$ (Klein et al. 2021). Hence, it is reasonable to expect that a meta-classifier may be able to "learn" (model) how to properly combine information from those methods (base classifiers), in order to get better performance than of each method individually.

In this paper, we applied a machine learning algorithm for this meta-classification task (unprecedented in geodesy). To act as the predictive meta-classifier, we choose a Multilayer Perceptron (MLP), a neural network with at least one hidden layer that tends to perform well in a wide variety of applications (Faceli et al. 2011). Features of our created database were composed of predictions from IDS, WML1 and SML1, and residuals of adjustment procedures that are part of these methods. Promising results pointed a higher accuracy of our proposed MLP-based meta-classifier.

2. Materials and Methods

In this work, we studied the leveling network configuration of Figure 1 (adapted from Ghilani 2010), with $m=14$ independent observations (height differences), $n=6$ unknowns (station heights) and one fixed control station A of known height. In the ascending order of observations index, the standard deviation of observations σ_i (in *meters*) were: [0.018, 0.019, 0.016, 0.021, 0.017, 0.021, 0.018, 0.022, 0.022, 0.021, 0.017, 0.020, 0.018, 0.020].

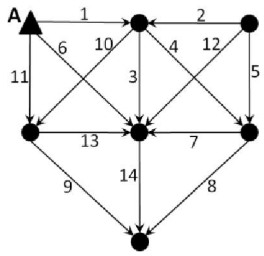

Figure 1. Configuration of the leveling network (adapted from Ghilani 2010)

In order to perform the experiments, we followed the workflow of Figure 2. It consists of four main stages: dataset creation, feature selection, meta-classifier training and evaluation. Python codes of the experiments were written and executed in (web-based) Google Colaboratory notebooks.

Figure 2. Workflow of the experiments

Our dataset was created based on Monte Carlo (MC) simulation. At first, given its matrices **A** and $\mathbf{\Sigma_l}$, we simulated different MC scenarios for the analyzed geodetic network. Random errors of each observation e_i, $i=(1,2,\ldots,m=14)$ were generated according to a multivariate normal distribution $\mathbf{e} \sim N(0, \mathbf{\Sigma_L})$. In scenarios with outlier, we added a gross error (with random sign) to the random error of respective randomly chosen observation. Magnitudes of gross errors were generated (with uniform distribution) in intervals from $3\sigma_i$ to $6\sigma_i$ ($3\text{-}6\sigma_i$) and from $6\sigma_i$ to $9\sigma_i$ ($6\text{-}9\sigma_i$).

For each MC scenario we generated the following 6 groups of $m=14$ features (each group with the same type of feature for each of the 14 observations of analyzed geodetic network), totaling 84 features. In binary classifications by base classifiers (IDS, WML1 and SML1) we adopted the values "1" if outlier and "0" otherwise. We adopted critical values $w_{IDS}=w_{WML1}=w_{SML1}=3.29$, as it is a common choice in the literature (Suraci et al. 2021). Being $i=(1,2,\ldots,m=14)$, the 6 groups of features were:

1) $|v_i|_{LS}$: absolute LS residual of each observation.
2) $|v_i|_{WML1}$: absolute WML1 residual of each observation.
3) $|v_i|_{SML1}$: absolute SML1 residual of each observation.
4) $(IDS\_result)_i$: binary classification of each observation in IDS procedure.
5) $(WML1\_result)_i$: binary classification of each observation in WML1.
6) $(SML1\_result)_i$: binary classification of each observation in SML1.

Regarding the labels, the dataset was created with 14 target attributes, each one representing a binary classification ("1" if outlier, "0" otherwise) of i^{th} observation. Hence, we have in this paper a multi-label classification supervised learning problem.

We divided our dataset in two partitions: Partition 1 (for feature selection and meta-classifier training stages) and Partition 2 (for performance evaluation stage). For Partition 1, the magnitudes of gross errors were generated (with uniform distribution) only in the *3-$6\sigma_i$* interval. We simulated 400,000 MC scenarios for Partition 1 (200,000 with no outliers and 200,000 with one outlier of magnitude *3-$6\sigma_i$*).

Note, however, that outliers in real geodetic networks may have magnitudes higher than *$6\sigma_i$*. Since it would not be viable to create a dataset based on an infinite range of outliers magnitude, we choose to consider only the "small" magnitudes (*3-$6\sigma_i$* interval) for Partition 1. Having our prediction model trained in such critical cases (closest cases to not having outlier at all), we expected that it would provide a good performance for the identification of "higher" outliers as well.

In this sense, for Partition 2 (evaluation) we also considered "higher" outliers. We simulated 150,000 new MC scenarios (50,000 without outliers, 50,000 with one outlier in *3-$6\sigma_i$* interval and 50,000 with one outlier in *6-$9\sigma_i$*). Table 1 summarizes MC scenarios of each partition.

Table 1. Quantity of MC scenarios for dataset creation

	0 outliers	**1 outlier *3-$6\sigma_i$***	**1 outlier *6-$9\sigma_i$***
Partition 1	200,000	200,000	xxx
Partition 2	50,000	50,000	50,000

In order to properly reduce the dataset dimensional space, but retaining meaningful characteristics of the original data, dimensionality reduction can be performed by feature selection or feature projection techniques. Although the latter generally has lower computational cost, it does not preserve data semantic. Therefore, since we intended to also check the most relevant features of our original dataset, we have chosen a feature selection technique for the dimensionality reduction.

We adapted the forward feature selection (FFS) technique (Faceli et al. 2011) with a new approach herein called "FFS in groups". As by the "standard" FFS, we have begun feature selection with an empty set of selected features. However, here we are calling it "in groups" because in the first iteration, instead of generating one predictive model for each of the 84 features, we generated 6 predictive models, each of them taking into consideration one of the 6 groups mentioned with respective 14 features. Then, we build reduced datasets considering combinations of two different groups of features (instead of two different features) and so on until no new combination of groups showed improvement in the model accuracy. Hence, our "FFS in groups" approach is much less computationally expensive than "standard" FFS.

For "FFS in groups", we randomly selected 300,000 instances for training and 100,000 for validation (in Partition 1). Predictive model was composed of a fully-connected MLP with one hidden layer of 100 neurons and learning rate of 0.001. As a result, the selected groups of features were $(\text{IDS\_result})_i$ and $|v_i|_{SML1}$ (totaling 28 features). Thereafter, considering only these 28 features, we used all 400,000 instances (of Partition 1) to train the meta-classifier with the same mentioned architecture.

Finally, we evaluated (using Partition 2) our MLP-based meta-classifier against IDS, WML1 and SML1. In order to provide a fair comparison among all classifiers, the evaluation considered the same false positive rate α (the rate of scenarios with no outliers, but in which at least one outlier was identified by respective classifier) for all of them. At first, we computed α for our MLP previously constructed by testing it in the 50,000 MC scenarios with no outliers. Thus, we computed the w_{IDS} that provides the same α for IDS with the procedure of Klein et al. (2021); and, we computed the w_{WML1} and w_{SML1} that provide same α in WML1 and SML1, respectively, by the procedure of Suraci et al. (2021).

3. Results and Discussion

In MLP evaluation, we obtained α=5.09%. Then, considering this same α, we computed w_{IDS}=2.87422, w_{WML1}=3.91034 and w_{SML1}=3.98099, and performed the evaluation of IDS, WML1 and SML1 with these critical values, respectively. Table 2 shows the accuracy and the 95% confidence interval of each classifier in evaluation experiments.

Table 2. Accuracy of the MLP meta-classifier, IDS, WML1 and SML1

		MLP	**IDS**	**WML1**	**SML1**
0 outliers (no false alarm):		**94.91%** (±0.21%)	**94.91%** (±0.21%)	**94.91%** (±0.21%)	**94.91%** (±0.21%)
1 outlier	$3\text{-}6\sigma_i$	**63.59%** (±0.42%)	61.65% (±0.43%)	52.30% (±0.44%)	54.25% (±0.44%)
	$6\text{-}9\sigma_i$	**97.95%** (±0.12%)	94.13% (±0.21%)	90.08% (±0.26%)	91.20% (±0.25%)

Firstly, we can verify that the MLP had the highest accuracy for both magnitude intervals of the outlier ($3\text{-}6\sigma_i$ and $6\text{-}9\sigma_i$). This result proves that for the analyzed geodetic network and considering scenarios without outliers or with one outlier, it was possible to obtain higher accuracy with a machine learning-based meta-classifier.

Moreover, even though we trained the MLP with only $3\text{-}6\sigma_i$ instances, we can see that the difference of accuracy from MLP to IDS (the second best) was even most significant in $6\text{-}9\sigma_i$ interval, which suggests that this was a good strategy for training. In fact, in $6\text{-}9\sigma_i$ interval the MLP failed less than half the times of IDS.

Now comparing only the classifiers from the main approaches for outlier identification in geodetic networks, we can see that the IDS performed the best. However, it is interesting to note that, although being a simplification of WML1, SML1 had best accuracy between the two of them.

4. Conclusions and future work

In this work, we successfully applied meta-classification to the identification of outliers in a leveling network. Promising results, in addition to many relevant points raised, highlight the potential of this research and justify its continuity.

We presented a new approach to FFS, herein called “FFS in groups”, which may be a valuable heuristic for geodetic networks and needs further investigations. Although there is no guarantee that our “FFS in groups” reaches exactly the optimum feature selection, it makes the FFS much less computationally expensive.

It is remarkable that $|v_i|_{SML1}$ group was selected (together with $(IDS\_result)_i$) in feature selection, while no groups related to WML1 was selected, which means that only SML1 was able to improve IDS accuracy. Besides, in the evaluation of all classifiers, SML1 presented better performance than WML1. These results emphasize the need for more investigations about the SML1 for outlier identification, something already mentioned by (Suraci et al. 2019).

In special, the proposed MLP-based meta-classifier presented the highest accuracy among all classifiers. This suggests a promising potential for the development of a new approach for outlier identification in geodetic networks, based on meta-classification with machine learning algorithms.

Future works shall optimize MLP hyperparameters; test other learning models performing the meta-classifier; try other data granularity approaches for dataset creation (e.g., one geodetic observation per line, instead of one geodetic network); and consider features with residuals from other adjustment procedures, such as Minimum L∞-norm, and predictions from other base classifiers originally from Geodesy, as the Sequential Likelihood Ratio Tests for Multiple Outliers (SLRTMO) procedure (Klein et al., 2017).

Besides, we trained our meta-classifier with scenarios of no outliers and "small" outliers ($3\text{-}6\sigma_i$ interval). Even though this seemed as a good strategy, other approaches for training must be considered, mainly for multiple outliers. In fact, experiments of this paper must be extended for multiple outlier scenarios.

Finally, meta-classification for outlier identification should be applied to other types of geodetic networks, such as Global Navigation Satellite Systems (GNSS) networks. In addition, geodetic networks with different redundancy must be considered, as this factor has significant influence in the accuracy of outlier identification.

References

Amiri-Simkooei, A. R. (2018). On the use of two L1 norm minimization methods in geodetic networks. *Earth Observ. Geomat. Eng.*, 2(1), 1-8.

Baarda, W. (1968). *A testing procedure for use in geodetic networks*. Publications on Geodesy, New Series, v. 2, n. 5. Netherlands Geodetic Commission.

Faceli, K., Lorena, A. C., Gama, J. and Carvalho, A. C. P. L. F. (2011). *Inteligência Artificial: uma abordagem de Aprendizado de Máquina*. LTC.

Ghilani, C. D. (2010). *Adjustment Computations: Spatial Data Analysis. 5th. edn*. John Wiley & Sons.

Goldschmidt, R., Passos, E., Bezerra, E. (2015). *Data mining: Conceitos, técnicas, algoritmos, orientações e aplicações. 2nd. edn*. LTC.

Klein, I., Matsuoka, M. T., Guzatto, M. P., and Nievinski, F. G. (2017). An approach to identify multiple outliers based on sequential likelihood ratio tests. *Survey review*, 49(357): 449-457.

Klein, I., Suraci, S. S., Oliveira, L. C., Rofatto, V. F., Matsuoka, M. T. and Baselga, S. (2021). An attempt to analyse Iterative Data Snooping and L1-norm based on Monte Carlo simulation in the context of leveling networks. *Survey Review*. doi: 10.1080/00396265.2021.1878338.

Suraci, S. S., Oliveira, L. C. and Klein, I. (2019). Two aspects on L1-norm adjustment of leveling networks. *Revista Brasileira de Cartografia*, 71(2): 486-500.

Suraci, S. S., Oliveira, L. C., Klein, I., Rofatto, V. F., Matsuoka, M. T. and Baselga, S. (2021). Monte Carlo-Based Covariance Matrix of Residuals and Critical Values in Minimum L1-Norm. *Mathematical Problems in Engineering*. doi: 10.1155/2021/8123493.

Teunissen, P. J. G. (2006). *Testing Theory: an introduction. 2nd. edn*. Delft University Press.