2nd URSI AT-RASC, Gran Canaria, 28 May – 1 June 2018

On development of the forecasting model of GNSS positioning performance degradation due to space weather and ionospheric conditions

Mia Filić(1)
(1) Faculty of Engineering, University of Rijeka, Rijeka, Croatia, e-mail: filicmia@gmail.com

Abstract

Space weather and ionospheric conditions effects on the Global Satellite Navigation System (GNSS) positioning performance and operation have already been identified. However, the qualification of this relationship is still a subject of scientific activities. A model forecasting the level of GNSS positioning performance degradation caused by space weather and ionospheric dynamics represents a valuable scientific goal. This manuscript addresses the refinement in forecasting model development procedure achieved through utilisation of selected supervised machine learning method based on Linear Models (LM) and Component Analysis (PCA) on experimentally collected data set of the quiet space-weather period.

1. Introduction

Studies indicating relations between space weather and ionospheric conditions and GNSS are cross-validated and accepted. However, the characterisation of the relationship is still the subject of scientific scrutiny. A model explaining relation of space weather effects on GNSS positioning performance would make a great supplement to GNSS-based applications and assist operators and users providing the evidence-based GNSS positioning performance degradation alerts [3], [4]. The alerts can then be additionally used to support the preparation of related risk mitigation operations [3]. This paper addresses the problem, establishes the methodology, and details development and validation of several models of space weather and ionospheric effects on GNSS positioning performance.

This manuscript is structured as follows. Section 2 provides description of the data set. Methods and methodology for model development are explained in chapter 3 and 4, respectively. Section 5 explains results which are summed up in chapter 6 together with manuscript's conclusions.

2. Data description

A data set was created combining experimental data from internet-based archives and observations at Faculty of Maritime Studies, University of Rijeka, Croatia for day in June, 2007 (DOY167 in 2007). Each variable in data set is described in Table 1. All variables are numerical.

Table 1 Related indices in the assembled data set

Index	**Parameter description**	**Remarks**
DOY	Timestamp	Serves for indication
Bx	Earth's magnetic field density - x component	Geomagnetic index, taken from the US NOAA archive
By	Earth's magnetic field density - y component	As stated above
Bz	Earth's magnetic field density - z component	As stated above
SFD	Solar flux density	Space weather index, taken from the US NOAA archive
SSN	Sunspot number	As stated above
Ap	Planetary geomagnetic A index	Geomagnetic index, taken from the US NOAA archive
Kp	Planetary geomagnetic K index	As stated above
Dst	Disturbance storm time index	As stated above
f0F2	Critical freq. of F2 layer	Ionospheric index, taken from the US NOAA archive
f0Es	Critical freq. of Es layer	As stated above
SID	Reference station signal strength observed with Sudden Ionospheric Disturbance (SID) monitor	Ionospheric index, observation collected at Faculty of Maritime Studies, University of Rijeka, Croatia

TEC	Total Electron Content (as derived from dual -frequency GNSS pseudorange observations)	Ionospheric index, derived from dual-frequency GNSS pseudorange data collected at IGS Matera, Italy reference station
sdTEC	Standard deviation of TEC	As stated above
d_fi	Observed equivalent positioning error (latitude) [m]	GPS, positioning performance index,derived from dual-frequency GNSS pseudorange data collected at IGS Matera, Italy reference station
d_lam	Observed equivalent positioning error (longitude) [m]	As stated above
d_plane	Observed equivalent positioning error (horizontal) [m]	As stated above

3. Methods

3.1 Linear regression model

With linear regression model [1], one models the relationship between *dependant* and *predictor* variables. Linear model with n *predictor* variables has a form:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n + \varepsilon \qquad (1)$$

where y is *dependant* or *response* and $x_1, x_2, ..., x_n$ are *independent* variables or *predictors*. $\beta_0, \beta_1, ..., \beta_n$ are regression coefficients. The random variable ε is an error term. It represents random fluctuations of a model. In linear modeling, one often assume normality of the error term and statistical independence of the error term from the response. The model development procedure requires the estimation of regression coefficients which is usually done using experimentally observed (n+1)-tuples of values of $x_1, x_2, ..., x_n$ and y. The estimated model can than be used to predict or forecast y for an observed n-tuple $x_1, x_2, ..., x_n$ and to estimate the accuracy of the prediction. Above the prediction and forecast, a linear model can also be used to summarise or explain observed data. When fitting linear model to observed data set, we use supervised learning [1]. Coefficients of the regression models are determined in the manner so the model becomes optimal in the sense of minimisation of the least square errors.

3.2 Principal component analysis

Principal component analysis (PCA) [2] is a statistical technique which transforms a data set with possibly linearly correlated *predictor* variables into a data set of linearly uncorrelated variables called Principal Components. The number of principal components is less or equal to the number of the initial input variables. The first k principal components comprise the most of the variability of a data under scrutiny among all competing transforms addressing the initial data set reduced to k variables. Consequently, the PCA is commonly used to reduce the dimensionality of a data set thus reducing the complexity of the model reducing: (1) the number of degrees of freedom of the hypothesis thus reduces the risk of overfitting, (2) the computable power of the algorithm. Reduction of dimensionality can also provide better visualization, insight into a data and a model outlook.

4. Methodology

Opposing the methodology in [4], in this study, only one ML model type (method) is considered. A selected model type is based on (multiple) linear regression. The aim is to identify the optimal model of the type (for the matter) through expansion of the machine learning-based model development process described in [4].

A multiple linear model is trained on 70% of the data. At the same time, t-test is run with respect to $H_0 : \beta_i = 0$ hypothesis for each *predictor* variable x_i $i \in \{1, 2, ..., k\}$, where k is a number of different *predictor* variables. As null-hypothesis is rejected when p-value is small, a small p-value indicates the predictor variable is likely to contribute to the model's ability to describe the process under observation. For a closer assessment of the trained model's quality, the ANOVA F-test is utilised with the respect to the null-hypothesis defined as: H_0 : *x_i addition in the model having the same response variable and all other predictors makes a significant change to the model.*
Again, rejecting the null-hypothesis marks selected predictor variable as likely to be an important (contribution) to the model. Lastly, residuals and forecast quality of the model are assessed (on the remaining data) using quantile-quantile (QQ) diagram and selected goodness-of-fit statistic, R-squared (R2) statistic, respectfully. R-squared represents the percentage of the response variable variation which is explained by a linear model and is graphically illustrated plotting predicted-observed diagram. Moreover, R2 provides an estimate of the relation between model and response strengthness which is formally proved utilising F-test.

For this study, three different linear models with the same response variable are observed. The first model (model A) is a linear regression model which utilises all data set variables, apart form d_fi, d_lam, d_plane. The second model (model B) is trained after the analysis of the t- and F test p-values of the model and transforming the predictor variables set. A new predictor variables set contains only predictors which have both, t- and F-test

p-values, valued less than 0.05. To obtain model C predictor variable set, model B predictor variable set is rotated utilising PCA expressing the values data set in Principal Components (PC). All three models are tested for quality and compared to the previous one. We label a model as of poor quality if its R2 value is less than 0.6, of good quality if its R2 value is between 0.6 and 0.8 and has at least normal residuals, and of very good quality if its R2 value is greater than 0.8 and has at least normal residuals and all predictors stated as relevant.

Comparing to the previous study [4], the initial predictor variables set is established in the same manner, while the response is different, d_plane is replaced with d_lam. As d_plane is directly derived from d_fi and d_lam variable, one forecasting d_fi and d_lam can forecast d_plane. The methodology is applied in the open-source programming environment for statistical computing R [1], [5].

5. Results

5.1 Linear modeling

Evaluation of the model A revealed that all t-test p-values are valued less than 0.05, thus indicating the significant impact on the predictor variable (in a sense of linear modeling) of all variables in predictor variable set. Furthermore, SFD and SSN variables show lack of variability by being constant in observed data set, thus being obsolete for the model development. On the contrary, F-test indicates that several variables (Bx, By and SID) can not be claimed to be significant addition to the model, since the respective F-test p-values associated with three variables exceed any (reasonable) significance level (0.1, 0.05, and 0.01).

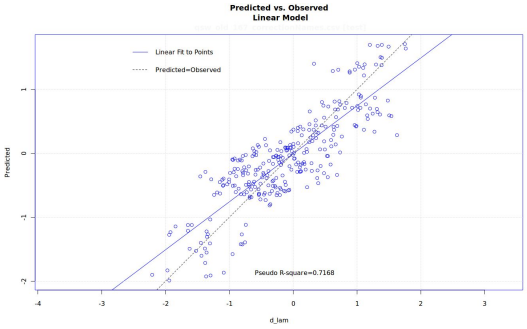

Figure 1 Model A: Predicted-observed diagram

Predicted-observed and QQ diagrams, presented in Figure 1 and 2, classifies model A as of good (forecast) quality according to criteria explained in Section 4. The analysis of the QQ diagram does not provide the evidence for rejection of the normality of residuals hypothesis and R2 is greater than 0.7.

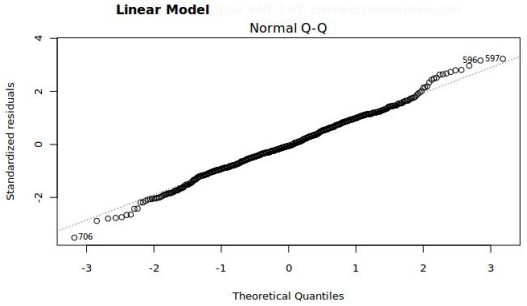

Figure 2 Model A: QQ diagram

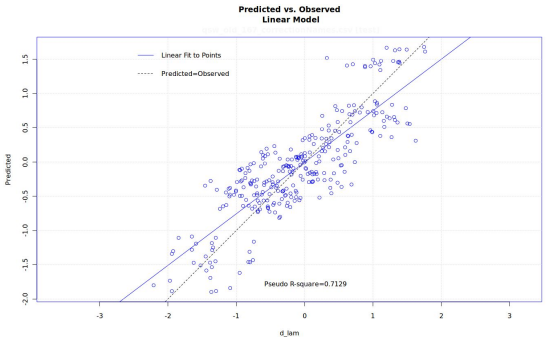

Figure 3 Model B: Predicted-observed diagram

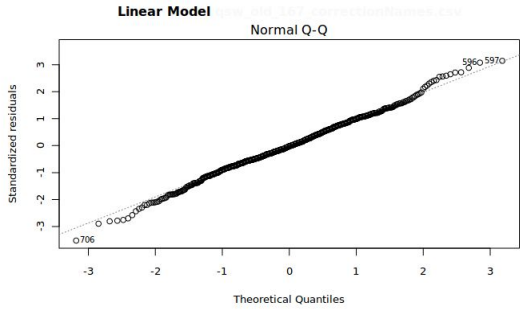

Figure 4 Model B: QQ diagram

In development procedure of the model B, the set of predictor variables should comprise only those predictors identified as significant for the model B by both F- and t-tests. In other words, one seeks to meet the linearity condition for the model. In order to fulfill the requirement, either the initial variables data set is reduced or selected variables are transformed. As reduction of the initial variables set results in significant reduction in quality, the model B predictor variables set is established transforming the initial variables set as follows. Bx, By, Bz are recentered, f0F2 and f0Es re-scaled using the logarithmic function, and SID is transformed in a way to obtain zero median with absolute value standard deviation equal to one. Both the F- and t-test p-values were valued less than 0.05, thus confirming the success of the model B training process. The predicted-observed diagram at Figure 3 and QQ diagram at Figure 4 show no significant reduction in model forecasting quality. Model B is of the forecasting quality comparable to model A, taken as the

control one, but without statistically insignificant variables that preserve the unwanted noise.

5.2 Principal component analysis

The approach taken in the model C development aimed at model complexity reduction using the PCA. First, the linear model is trained on the model B data set. Then, the linear model's F- and t- tests p-values are assessed. The fifth PC t-test p-value of 0.656 exceeds significantly the referent value of 0.05 which renders this predictor variable insignificant for the model development. Finally, the linear model is trained again but on the set of predictor variables without the fifth PC. The resulting model C has at least sustained (forecasting) quality comparing to the model A, as evident from (forecasting) quality results depieced in Figure 5 (Predicted-observed diagram) and 6 (QQ diagram). The reduction in number of predictors reduces the model complexity which makes model C computationally more efficient than model A.

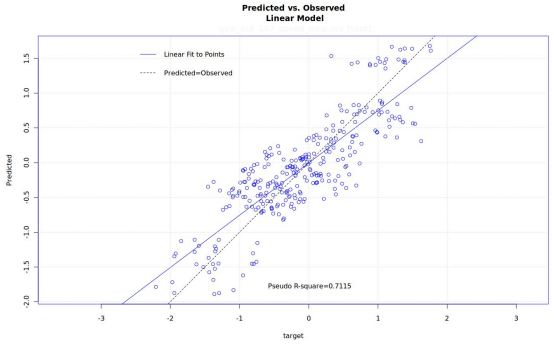

Figure 5 Model C: Predicted-observed diagram

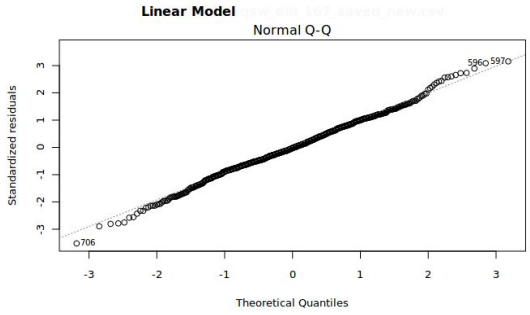

Figure 6 Model C: QQ diagram

A closer consideration of quality assessment of the developed models reveals that the model C yields the experimental distribution of residuals that is the closest to the equivalent normal distribution. All together indicates the model C to be the best candidate for the forecasting model in focus.

6. Discussion and conclusion

This manuscript presents the results of the study on model development approaches in forecasting the GNSS positioning performance dependence on space weather and ionospheric conditions. Three linear models were developed based on experimentally collected data sets using separate model development approaches. In continuation of the previous research [4] and [5], the improvements in the model development process were demonstrated. Improvements are based on optimisation of the process through identification of statistically significant predictors and achievements of the computational efficiency reduction utilising PCA. Optimisation was achieved at no cost for the model forecasting quality and remained at requested level for all three modelling approaches (R2 value valuing a bit higher than 0.7 and not rejecting normality of residuals).

In future research, it is expected to focus on the characterisation of the GNSS positioning performance in other space weather and ionospheric conditions scenarios, such as those with mild disturbances or a large ionospheric storm and the determination of the appropriate methodology for statistical forecasting model development in the given GNSS positioning environment conditions.

7. Acknowledgements

Author appreciates the support for participation from the AT-RASC 2018 organisers and USRI.

8. References

1. E. Bradley and T. Hastie, *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science, Cambridge University Press. Cambridge, UK*, 2016.

2. A. C. Cadavid, J. K. Lawrence and A. Ruzmaikin, "Principal Components and Independent Component Analysis of Solar and Space Data," *arXiv.org pre-print archive*, 2017, doi:https://doi.org/10.1007/978-0-387-98154-3_5.

3. P. Cannon, *Extreme Space Weather: impacts on engineered systems and infrastructure, The Royal Academy of Engineering. London, UK* , 10 May 2013.

4. M. Filić, "A comparative study of forecasting methods for space weather-caused GNSS positioning performance degradation, " Presentation given at UN/USA Workshop on ISWI, Boston College. Chestnut Hill. MA, August 2017.

5. R. Filjar, M. Filić and E. Mirmakhmudov, "Categorisation of space weather and GNSS positioning quality indices for estimation of GNSS positioning performance degradation - Accepted for publication in Proc of 11th Annual Baška GNSS Conference, Baška, Krk Island, Croatia," May 2017.