CIplus
Band 5/2018

Modelling Zero-inflated Rainfall Data through the Use of Gaussian Process and Bayesian Regression

Margarita Alejandra Rebolledo Coy and Thomas Bartz-Beielstein

Technology
Arts Sciences
TH Köln

Modelling Zero-inflated Rainfall Data through the Use of Gaussian Process and Bayesian Regression

Margarita Alejandra Rebolledo Coy and Thomas Bartz-Beielstein

Institute for Data Science, Engineering, and Analytics, TH-Köln

November 29, 2018

1 Introduction

Rainfall is a key parameter for understanding the water cycle. An accurate rainfall measurement helps in the development of more accurate hydrological models. These hydrological models can be used later in the design of better management plans for the available water resources or in the implementation of flood or drought warning systems for regions at risk. In the recent decades, rainfall estimation done by satellite products have been made available, providing a worldwide high spatio-temporal estimation of precipitation. However, as these satellite rainfall estimates (SRE) are done using indirect measurements from the satellites' sensors a validation process needs to be carried out in order to avoid their incorrect use [2]. Following [1] we aim to generate a bias-corrected estimate of rainfall using satellite data and rain gauge data. Rain gauges are rainfall sensors located in a network in some given area. One of the main obstacles in using these sensors as a reliably source of precipitation measurement is the lack of coverage in large areas. Using the available rain gauges we wan to calibrate the SREs for the point in which the rain gauge is located and its adjacent area. For this we use Guassian process regression and Bayesian linear regression on a rainfall data set.

2 Data Description

The selected rainfall data set comes from the Imperial basin located in Chile. This is a relatively small area unevenly covered with 13 rain gauges. One of the important characteristic this area presents is its pluvial hydrological regime, meaning most of its water comes from rainfall. The collected data covers a range of 13 years, from 2003 to 2015. The rainfall measurements are organised in 13 tables each with 4748 data points. All tables contain the following information:

- The date on which the measurement was taken.

- The precipitation value in millimetres (mm) measured by the rain gauge (observed values).
- The SRE precipitation value in mm aggregated yearly recorded for the specific station area (SRE_annual).
- The SRE precipitation value in mm aggregated seasonally recorded for the specific station area (SRE_seasonal).

The data in all of the 13 tables show very similar characteristics, with high dispersion and a lot of data points in or around zero, as illustrated in fig. 1.

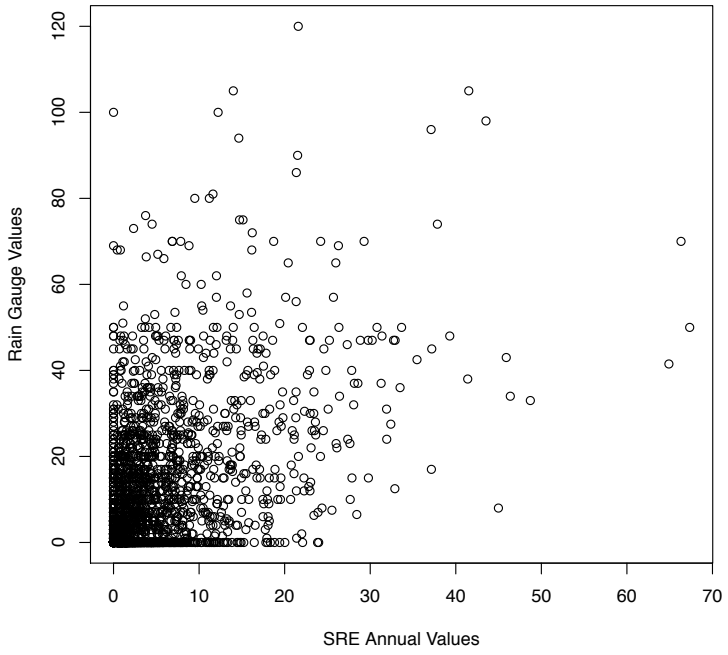

Figure 1: Visualization of one of the data tables. The observed values vs SRE_annual show a high dispersion with a lot of points concentrated around zero

3 Experiments

Two regression models are fitted, the first using the R package SPOT [3] for the Gaussian regression. The second implements Bayesian regression using the statistical language STAN [4]. We use Root Mean Square Error (RMSE) and Kling-Gupta Efficiency (KGE) to evaluate the goodness-of-fitness (GOF) of the models. As a baseline we define the RMSE and KGE that the observed values have against the measured SRE. If a model

surpasses the baseline GOF then it is considered that this model gives a better approximation of the rainfall than the raw SRE. To test for stability the regression models were run 10 times with different starting points in each of the data tables.

4 Results

Given the large amount of data points cluster Kriging was implemented when executing the Gaussian regression on the yearly SRE data. According to our results neither of the models gave a good approximation of rainfall when using the yearly data. On the other hand, Gaussian regression delivered a better approximation on the rainfall real values when using seasonal SRE data. In this point it was noted that Bayesian regression was not able to capture medium to heavy rainfall events.

5 Conclusion and Future work

Overall Gaussian process regression showed a better performance for this data in comparison to Bayesian regression. However with its high time complexity it may be a problem when applied to bigger data sets. Cluster Kriging was implemented as a solution to this problem however this increased the error in the model. In the case of the Bayesian regression it was noted that its posterior distribution was not able to escape a very reduced area, losing information of heavy rainfall events. In future works, we would like to explore a different approach to cluster Kriging that can reduce the amount of introduced error as well as different distributions and constraints for the Bayesian regression.

References

[1] M. Zambrano-Bigiarini, A. Nauditt, C. Birkel, K. Verbist, L. Ribbe. "Temporal and spatial evaluation of satellite-based rainfall estimates across the complex topographical and climatic gradients of Chile". In: *Hydrology and Earth System Sciences*. 21.2. 2017.

[2] M. Gebremichael, E.N. Anagnostou, M.M. Bitew. "Critical Steps for Continuing Advancement of Satellite Rainfall Applications for Surface Hydrology in the Nile River Basin". In: *jJAWRA Journal of The American Water Resources Assosiation* 46.2. 2010.

[3] T. Bartz-Beielstein, C. Lasarczyk, M. Preuss "Sequential Parameter Optimization". In: *IEEE Congress on evolutionary computation*. 2005.

[4] Stan Development Team. "RStan: the interface for Stan in R" Package version 2.16.2 *http://mc-stan.org* 2017

Kontakt/Impressum

Diese Veröffentlichungen erscheinen im Rahmen der Schriftenreihe "CIplus". Alle Veröffentlichungen dieser Reihe können unter
https://cos.bibl.th-koeln.de/home
abgerufen werden.

Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.
Datum der Veröffentlichung: 07.11.2018

Herausgeber / Editorship

Prof. Dr. Thomas Bartz-Beielstein,
Prof. Dr. Wolfgang Konen,
Prof. Dr. Boris Naujoks,
Prof. Dr. Horst Stenzel
Institute of Computer Science,
Faculty of Computer Science and Engineering Science,
TH Köln,
Steinmüllerallee 1,
51643 Gummersbach
url: www.ciplus-research.de

Schriftleitung und Ansprechpartner/ Contact editor's office

Prof. Dr. Thomas Bartz-Beielstein,
Institute of Computer Science,
Faculty of Computer Science and Engineering Science,
TH Köln,
Steinmüllerallee 1, 51643 Gummersbach
phone: +49 2261 8196 6391
url: http://www.spotseven.de
eMail: thomas.bartz-beielstein@th-koeln.de

ISSN (online) 2194-2870

Technology
Arts Sciences
TH Köln

Technology
Arts Sciences
TH Köln