A Web Processing Service for Validating Interpolation

J. de Jesus [1,*], R. Barillec [2], G. Dubois [1] and D. Cornford [2]

[1] European Commission, Joint Research Centre, Institute for Environment and Sustainability

jorge.de-jesus@jrc.ec.europa.eu, gregoire.dubois@jrc.ec.europa.eu

[2] NCRG / Aston University
barillrl@aston.ac.uk, d.cornford@aston.ac.uk

Abstract. An interoperable web processing service (WPS) for the automatic interpolation of environmental data has been developed in the frame of the INTAMAP project. In order to assess the performance of the interpolation method implemented, a validation WPS has also been developed. This validation WPS can be used to perform leave one out and *K*-fold cross validation: a full dataset is submitted and a range of validation statistics and diagnostic plots (e.g. histograms, variogram of residuals, mean errors) is received in return. This paper presents the architecture of the validation WPS and a case study is used to briefly illustrate its use in practice. We conclude with a discussion on the current limitations of the system and make proposals for further developments

1 Introduction

An interoperable framework for real time automatic mapping of critical environmental variables has been developed in the frame of the INTAMAP project (http://www.intamap.org). To better assess the quality of the maps produced by the interpolation server, a cross-validation service has been developed. Cross validation is generally used to determine how well a model is performing in predicting a data set. When performing a spatial interpolation, the data set is partitioned into one subset (called the training set) which is used for parameter estimation and interpolation and another one (called the validation set) for validation. Predicted values can so be contrasted against a set of true values. In *K*-fold cross validation, the input dataset is divided into *K* partitions. Of the *K* partitions, a single one is retained as the validation data for testing the spatial interpolation method, and the remaining *K*-1 partitions are used as training data. The cross-validation process is then repeated *K* times, with each of the partitions used once as the validation data. Leave-one-out cross validation (LOOCV) is a *K*-fold cross validation in which *K* equals the size of the input data set meaning that each point is removed in turn. We refer the reader to [4] for more details on these validation techniques.

[*] Corresponding author

2 SYSTEM ARCHITECTURE

INTAMAP's cross-validation service follows the SOA (Service Oriented Architecture) model, and uses OGC's Web Processing 1.0.0 as communication specification[1]. This implementation offers a high level of flexibility as it allows the user to test results generated by any interpolation service without having a direct access to it.

With the general *GetCapabilities*, *DescribeProcess* and *Execute* requests defined by the WPS the cross validation service is described and the requests executed, respectively. The service was built in a Linux-Apache-Python-R (LAPR) system using PyWPS 3.0.0 as the major API for WPS requests and R graphic output. The different inputs / outputs are in XML format and defined by the WPS protocol as either LiteralValues (simple numerical or string input) or ComplexData (XML structured input).

2.1 Cross-validation sequence

The cross validation service was designed to be independent of the interpolation service and to allow end-users to define their own parameters when choosing the size of the validation set, and is relevant to any future interpolation services developed. It is thus a special client to an interpolation service that generates a number of requests that can be proportional to the size of the dataset. The sequence of the data exchange between the cross-validation and the interpolation server is summarized below:

Parsing of the input data set and checking availability of the interpolation server

1. Parsing of the dataset submitted by the user;
2. Parsing of any other inputs that might have been submitted (*K*-fold, method type, interpolation service location);
3. Parsing of the interpolation service location (server + URL) (submitted or default value);
4. Availability checking of the service and server using *GetCapabilities;*

Obtaining model and parameters values (dummy interpolation)

5. Submission of the full dataset to the interpolation server and request for a "dummy interpolation";
6. Parameters and models used to interpolate the data automatically are obtained from the "dummy" interpolation request sent to the interpolation server;

Cross-validation

7. Split of the original data set into subsets for training and validation (interpolation) according to the *K*-fold value;
8. Interpolation request of each subset using the parameters and model obtained from the dummy request;
9. Statistical analysis of the residuals and graphical output using R

[1] http://www.opengeospatial.org/standards/wps

Response

10. Assembling the WPS response for the client of the cross validation service.

2.2 Input for the cross validation service

The cross validation service will run with a simple dataset submission, without any other parameter specification:

- ObservationCollection, Observation & Measurement[2] format (mandatory)
- *K* value (optional)
- Interpolation Server (optional)
- Interpolation Process name (optional)

The Interpolation Process name allows for the user to assign one of the five interpolation methods currently supported by the interpolation services: idw (inverse weighting distance), automap, psgp (projected sequential Gaussian processes), copula-based geostatistics or automatic (the server will determine which method is best from the last four methods depending mainly on computational time constraints defined by the end-user as well as on the normality of the data).

2.3 Statistical output from the cross validation service

The cross-validation server will return information about the residuals (observed values minus the predicted values). Typically, the service will return statistics of the residuals (Mean Error, Mean Absolute Error, Mean Square Error, Root Mean Square Error), a few plots showing the distribution of the errors (histograms), their correlation with the input values (correlation plots of observed against estimated values) and a variogram of the residuals which can be used to assess how well the interpolation server managed to extract the spatial features of the modelled phenomenon. The variogram is an important diagnostic tool and can suggest further exploration to assess the possible presence of an underlying trend in the data.

To illustrate the use of cross-validation service, we tested the default automatic interpolation method by submitting the 467 daily rainfall observations used in the SIC 97 exercise [3] to the service. The results obtained using the LOOCV approach are compared in Table 1 to the results from the participants of the SIC97 exercise. In this case, the interpolation server used copula-based geostatistics as the default method.

Without discussing in depth the results of INTAMAP's automatic interpolation server, one will realize that the default choice of the service was as good as, or better than, the best results obtained by the experts who participated to SIC97, for both the criteria analysed.

[2] http://www.opengeospatial.org/standards/om

Table 1. Interpolation results using LOOCV: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) obtained by experts and INTAMAP's interpolation server

Interpolator	RMSE	MAE
INTAMAP interpolation service	**45.9**	*32.4*
Participant with lowest RMSE obtained in SIC97	53.1	36.7
Participant with lowest MAE obtained in SIC97	62.0	**32.0**

2.4 Technical Output from the cross validation service

The cross-validation service also returns information on the errors encountered and the time needed to compute the cross-validations. For a set of 467 points, the leave one out cross-validation needed 25 minutes to be computed, of which around 8 minutes (= 500 s) were used to exchange the 467 data sets through the internet and make a WPS request . This time seems prohibitive when knowing that the same calculation would require a few seconds on a stand-alone computer. However, generating and parsing the XML contents of the WPS requests and responses are not considered to be the main obstacles to a quick validation in contrast to the number of WPS request as well as the computation time of the interpolation server. The cross-validation time can be significantly reduced by using smaller values of *K* for the *K*-fold cross validation.

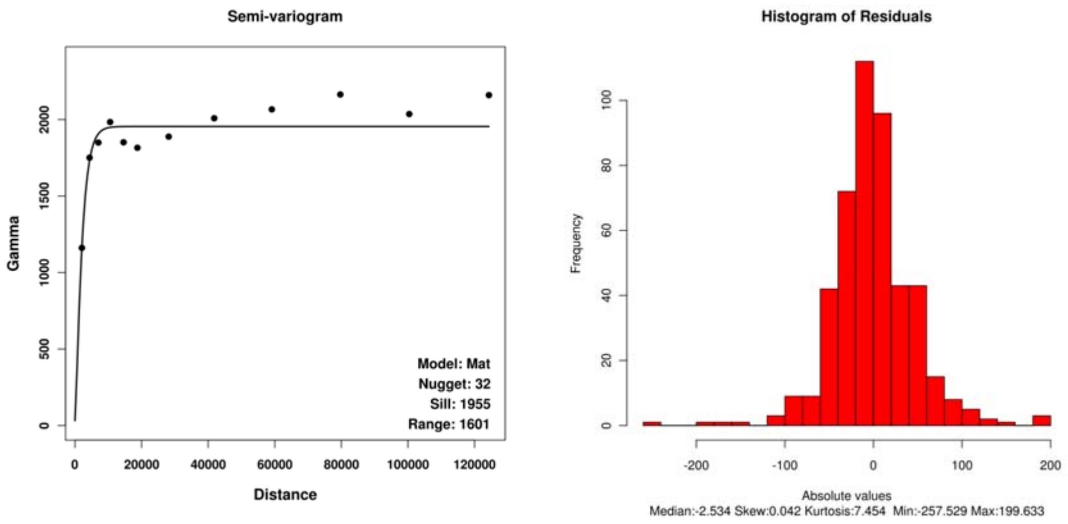

Figure 1: Variogram (left) and histogram (right) of the residuals obtained after cross-validation of the SIC97 data set using INTAMAP's interpolation server.

The graphical output presented by the cross-validation server is embedded in the WPS's response document as a SVG (Scalable Vector Graphic) format[3]. This format is an XML document that describes the graphic itself, therefore allowing for it to be opened directly in the web browser. An example of such graphics is shown in Figure 1 presenting the variogram, which still indicates a short scale spatial correlation in the residuals, and the histogram of the residuals.

3 DISCUSSION AND FURTHER DEVELOPMENTS

Validation can be seen from two perspectives. One, typically used in the computer science community, is the process of checking if something satisfies a certain criterion. Statistical validation is the process of assessing how well a model will generalise to locations without observations, and is crucial to model building, and often associated with the use of diagnostics to identify issues in models. The cross validation service might be used in both contexts: LOOCV can be used to assess whether one specific interpolation model fits the data set well enough (under a certain measure) to be used in an operational setting, while K-fold cross validation allows a more complete assessment of model fit and parameter uncertainty in the models (assuming parameter re-estimation).

As implemented LOOCV is simple (note some authors assume in LOOCV that model parameters are re-estimated, here they are treated as fixed) and probably not particularly suited to the web service (SOA) architecture employed, due to the need to repeatedly recalculate time consuming models for large data sets. For LOOCV we would in future recommend integrating LOOCV with fixed interpolation model parameters into a separate interpolation validation service, which could be accessed by the validation client with a single request. This service could then exploit many numeric tricks (e.g. sequential matrix inverse updates) internally to significantly speed up the process.

When implementing K-fold cross validation, where interpolation model parameters are recomputed for each training set, the web service framework has a much smaller communication overhead, and many attractive properties. In particular as future interpolation services are built (using a WPS interface) it will be simple to provide fast and reliable cross validation, together with diagnostic plots with very little additional coding.

A key benefit of using K-fold cross validation is that improved validation statistics and diagnostics can be computed. Given the difficulty in providing a unique and objective answer to the quality of interpolation results (see [2, 5] for discussions), the cross-validation service is currently only reporting back the residuals and their single point statistics, together with an informative variogram showing the 2 point statistics.

The INTAMAP interpolation service returns more than simply point predictions, it provides the full predictive distribution, or requested summaries thereof using UncertML [6]. This allows the assessment of the probabilistic predictions. In future im-

[3] http://www.svg.org/

plementations we will extend the validation metrics to include Receiver Operating Characteristic (ROC) curves, and the Mahalanobis distance, and relevant decompositions of the associated Mahalanobis errors [1] to provide improved diagnostics for kriging predictions which will allow more detailed assessments of the fitted model.

4 Acknowledgements

This work is funded by the European Commission, under the Sixth Framework Programme, by the Contract N. 033811 with the DG INFSO, action Line IST-2005-2.5.12 ICT for Environmental Risk Management. The views expressed herein are those of the authors and are not necessarily those of the European Commission. More information on INTAMAP can be found at http://www.intamap.org.

5 References

[1] L. S. Bastos and A. O'Hagan. Diagnostics for Gaussian process emulators. Research Report No. 574/08, Department of Probability and Statistics, University of Sheffield. Submitted to *Technometrics*. 2008

[2] D. Cornford. Why comparison studies are a waste of time: SIC2004 examined. In:, *Automatic mapping algorithms for routine and emergency monitoring data. Report on the Spatial Interpolation Comparison (SIC2004) exercise.* G. Dubois (Ed.), Office for Official Publications of the European Communities, Luxembourg; EUR 21595 EN, EC, pp. 61-68, 2005.

[3] G. Dubois. Spatial Interpolation Comparison 97. Introduction, and description of the data set. In: *Mapping radioactivity in the environment. Spatial Interpolation Comparison 1997*. In: G. Dubois, J. Malczewski & M. De Cort (Eds.), Office for Official Publications of the European Communities, Luxembourg; EUR 20667 EN, EC, pp. 39-44, 2003.

[4] B. Efron and G. Gong. A leisurely look at the bootstrap, the jackknife and cross-validation. *The American Statistician*, **37**(1): 36-48, 1983.

[5] K. G. van den Boogaart. The comparison of one click mapping procedures for emergencies. In: *Automatic mapping algorithms for routine and emergency monitoring data. Report on the Spatial Interpolation Comparison (SIC2004) exercise*. G. Dubois (Ed.), Office for Official Publications of the European Communities, Luxembourg; EUR 21595 EN, EC pp. 71-78, 2005.

[6] M. Williams, D. Cornford and L. Bastin. Describing and Communicating Uncertainty within the Semantic Web, In: *Uncertainty Reasoning for the Semantic Web Workshop, 7th International Semantic Web Conference*, 26 October Karlsruhe, Germany, 2008.