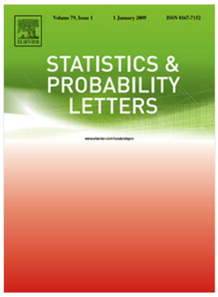
Accepted Manuscript



Statistical issues in radiosonde observation of atmospheric temperature and humidity profiles

A. Fassò, F. Finazzi, F. Madonna

PII: S0167-7152(18)30072-5
DOI: https://doi.org/10.1016/j.spl.2018.02.027
Reference: STAPRO 8149

To appear in: *Statistics and Probability Letters*

Statistical issues in radiosonde observation of atmospheric temperature and humidity profiles

A.Fassò[a], F.Finazzi[a], F. Madonna[b]

[a] University of Bergamo, Department MIPE, Viale Marconi 5, Dalmine, BG, Italy
[b] CNR-IMAA, C.da S. Loja, Tito Scalo, PZ, Italy

Abstract

Measurement uncertainty of atmospheric profiles obtained by radiosoundings is crucial in climate change studies. This paper shows how the understanding of geographic gaps of radiosonde networks calls for a functional approach able to handle spatio-temporal profile data, and related complexity issues are addressed.

Keywords: Functional data, spatio-temporal models, GRUAN, RAOB

1. Introduction

Measurement system uncertainty in climate data records (CDRs) and its impact in climate change assessment has been raised by various climate scientists and metrologists. This is especially true for observations of temperature and humidity provided by radiosoundings. Addressing this issue, the GRUAN reference measurement network (GCOS Reference Upper-Air Network, www.gruan.org) has been established, and started to provide valuable contribution to the understanding of measurement uncertainty, see Bodeker et al., 2016. Although GRUAN gives fully traceable measurements, its geo-

Email address: alessandro.fasso@unibg.it (A.Fassò)

graphic and historical coverage is quite limited. For this reason climatological studies are largely based on baseline measurement networks, which have an intermediate metrological quality but a larger spatio-temporal coverage.

In perspective, integrated data sets will include both ground based and satellite observations of land, sea and atmosphere. Moreover ensembles, obtained by various simulation techniques, are of increasing importance to describe uncertainty. Hence, considering the data growth rate we are facing, the size of such data sets is fast increasing from the order of terabytes to petabytes. This frame requires a new cooperation effort to elaborate new multi-discipli nary approaches and services for different types of users. Hence, the integration of atmospheric, metrology, statistical and computer sciences is considered a fruitful route to fully exploit the available historical CDRs of several Essential Climate Variables (ECVs) collected by satellite observations platforms and by ground based networks operating at the global scale.

In this frame the uncertainty of baseline networks is a challenging issue and the present paper aims at being a step to face this problem. In particular geographic gaps of temperature and humidity radiosonde networks are discussed and the need for advanced statistical methods is illustrated.

The rest of the paper is organized as follows. In Section 2 ongoing projects involving production and analysis of climate data sets in general and in particular radiosoundings are discussed and showed to call for new statistical developments. Section 3 considers spatio-temporal modeling for functional data in connection to network gap identification. Moreover Section 4 deepens some computational issues related to data and model size.

2. Data sets and projects

Copernicus is the European union's Earth observation programme (www.-coopernicus.eu). Its aim is to help in understanding how our planet and its climate are changing, the role played by human activities in these changes and how these will influence our daily lives. To do this Copernicus is involved in a complex set of systems, which collect data from multiple sources: earth observation satellites and in situ sensors such as ground stations, airborne and sea-borne sensors. It processes these data and provides users with reliable and up-to-date information through a set of services related to environmental and security issues.

In particular, the Copernicus Climate Change Service (C3S), which is operated by the European Centre for Medium-range Weather Forecasts (ECMWF, www.ecmwf.int), will provide comprehensive climate information covering a wide range of components of the Earth-system and timescales spanning decades to centuries. It will maximise the use of past, current and future earth observations (from in-situ and satellite observing systems) in conjunction with modeling, supercomputing and networking capabilities. This will produce a consistent, comprehensive and credible description of the past, current and future climate.

Various other projects developed in the frame of the Horizon 2020 program have been the scientific precursors of C3S. In particular Fiduceo (Fidelity and Uncertainty in climate data records from Earth Observations, www.fiduceo.eu) aims at bringing insights from metrology to the observation of Earth's climate from space (Merchand et al. 2017). Moreover, GAIA-CLIM (Gap Analysis for Integrated Atmospheric ECV CLImate Monitoring, www.gaia-clim.eu) aims at understanding gaps in integrated monitoring of upper troposphere and to improve our ability to use ground-based and sub-orbital observations to characterise satellite observations for a number of atmospheric ECVs, see Thorne et al., 2017. In fact despite that satellite Earth observation technology has undoubtedly facilitated the development of global climate change research, ground-based networks such as radiosonde networks are required to identify biases and issues in the satellite CDRs. Therefore, radiosonde observations remain an essential component of the observing system of systems. In this frame, geographic gaps are characterized by poor spatial coverage of a monitoring network.

On their turn, to represent a reliable and effective reference information all of these conventional anchor data sources must be harmonized and homogenized to achieve physical consistency of the decadal time series. Homogeneization, is essentially change detection and adjustment of data for any kind of known and quantifiable inhomogeneities (bias, change of sensors, calibration drift, local environment changes etc). Homogeneization methods have been developed for radiosonde has a long history, see for example Haimberger et al. (2012), Thorne et al. (2011) and Sherwood et al. (2008) . Although the statistical interpretation of these methods is very interesting, is omitted here for brevity. Harmonization is involved with the traceable characterization of the total uncertainty budget. For example the harmonization of temperature and humidity CDRs is one of the funded activity by C3S under contract C3S_311a_Lot3 (Madonna et al., 2017), and may benefit from the geographic gap analysis discussed in this paper.

3. Statistical issues and modeling

One of the objectives of GAIA-CLIM project is to understand the information content of ground based monitoring networks and to identify geographic gaps of these networks. We focus here on the network of the Universal RAwinsonde OBservation program or simply RAOB (www.raob.com), which has a global coverage with about 2400 stations and some decades with bi-daily prevailing temporal frequency. This type of baseline networks are also involved in the C3S harmonization problem.

From the statistical point of view, Fassò et al. 2014 and Ignaccolo et al., 2015, showed that atmospheric soundings may be conveniently described as functional data using appropriate basis function expansion, at least when these data can be handled as independent replications of the same model. In case of global networks spatial and temporal correlation must be considered. Various functional models have been considered for spatio-temporal data where usually time dynamics is embedded in the functional object, see e.g. Menafoglio et al. (2013) and Mateu and Romano (2017) and references therein for recent advances in the field.

The idea of spatially correlated functional data (e.g. Delicado et al., 2010, and Ruiz-Medina, 2012) may be extended, to handle a manifold domain such as the sphere and the temporal dimension (e.g. Porcu et al., 2016). Although the idea of modeling these data as spatio-temporally correlated functional data is quite natural, from the point of view of probability theory, this object may be considered as a stochastic process defined on *sphere* $\times$ *time* with values in a functional space. Alternatively, it may be considered as a stochastic process defined on a *spherical shell* $\times$ *time* with scalar values. In both cases, the full characterization of the underlying stochastic process is still under study, including the definition of flexible families of valid covariance functions (Porcu, 2017, private communication).

Taking into account computational burden and data dimensionality, the statistical model needs to be simple enough. Various solutions for modeling large spatial datasets have been proposed including nearest neighbor models (Vecchia, 2008, and Datta et al., 2016). A first step in computation reduction is to use separable models with discrete time, possibly after adjusting for relevant trends. Following this approach, maximum likelihood estimation can be based on the EM algorithm extending the multivariate dynamic coregionalization model (Finazzi and Fassò, 2014 and Calculli et al., 2011). This can be easily done by applying a multivariate spatio-temporal model to the coefficients of the basis function expansion. In some sense this approach is close to kriging for function-valued data discussed by Delicado et al. (2010). A relevant difference is related to the smoothing factor used to obtain the basis function coefficients: while in Delicado et al. (2010) a crossvalidation approach based on spatial prediction is used, for RAOB data, Fassò et al. (2017) proposed a criterion deploying metrological concepts, namely optimizing the approximation to the high quality GRUAN reference network data. Moreover this approach is easily coupled with extensions of block tapering as discussed in Section 4.

Using an appropriate spatio-temporal statistical model, a monitoring network geographic gap may be defined as a region where the uncertainty of the spatial forecast is larger than a threshold. The threshold may be based on statistical and/or metrological considerations. Since the variability of the spatial forecast error is influenced by the atmospheric variability, such a confounder may be controlled by adjusting for the effect of meteorology. A viable solution is to use the output of a numerical weather prediction model or a re-analysis such as ERA-Interim (Dee et al., 2011) as a model covariate, in fact these data are available with a reasonable resolution (about 80x80km grid) both on the RAOB stations for estimation and in the rest of the Earth for forecast.

4. High dimensionality challenges

The prediction problem of previous section is inherently high-dimensional due to the fact that the number q of coefficients per profile can be high, $q > 15$ say. In turn, it follows that the variance covariance matrices involved in model estimation are large, even when the number of stations is not very large. The problem becomes even bigger if multiple ECVs are jointly modeled to improve prediction, in which case the basis function coefficients add up.

Modeling spatio-temporal correlations across coefficients is not trivial, especially when the support is complex and the underlying process is non-stationary and anisotropic. In particular, the spatial correlation needs to be valid for the spherical shell and in the multivariate case. The problem becomes simpler if the statistical model includes a re-analysis model output as discussed in the previous section. In this case, stationarity and a reduced spatio-temporal correlation range are less strong assumptions.

Even when simple covariance functions are adopted, the estimation of the model parameters is computationally demanding in the multivariate case. For instance, a simple linear coregionalization model requires to estimate the parameter of a common correlation function and the elements of a correlation matrix the dimension of which is $q \times q$. This calls for efficient estimation methods when the model parameters are estimated using MCMC or the EM algorithm. In general, computational efficiency is attained allowing large matrices to be sparse without loosing unbiasedness and consistency of the estimators. Kaufmann et al., 2008 proposed covariance tapering. This approach allows to control the matrix sparsity but may have poor estimation properties as shown by Stein (2013), which suggested the simpler and better approach called block-tapering. In this case, the spatial locations are divided into blocks and each block contributes independently to the likelihood function. This allows to work with smaller matrices and to obtain consistent estimates speeding up the computation.

For spatio-temporal models with multivariate response, block-tapering may be extended from spatial blocks to spatio-temporal blocks and/or to spatio-temporal-response dimension blocks. As a result, in applications involving complexity similar to RAOB data set under consideration, computation time may be reduced by 10-20 times.

Although straightforward and easy to be implemented, the block-tapering approach still requires a more in-depth study about how it affects the estimation of the model parameters. Open questions includes the optimal definition of block sizes and block allocation. Additionally, attention must be paid when block-tapering is applied to the multivariate case. Matrices must be constructed in such a way that all the model parameters are identifiable when blocks are defined. Again, care must be taken when multiple ECVs are considered and the monitoring networks are unbalanced. Blocks must be defined in a way that the cross-correlation between ECVs can be consistently estimated. Finally, if the latent processes are non-stationary and anisotropic, the way block are defined may affect the estimation of the model parameters that control the nonstationarity and/or the anisotropy.

In our opinion, block-tapering is appealing since it does not require to alter the definition of the latent processes and the likelihood function simply factorizes across the blocks, allowing model estimation to be accomplished faster. On the other hand, effort must be spent to carefully defines blocks, possibly in an adaptive manner during the parameter estimation.

Acknowledgements

This research is partially funded by:
C3S_311a_Lot3 “Baseline And Reference Observations Network” Service Contract 1, under Copernicus Climate Change Service (C3S) Framework Agreement MWF/COPERNICUS/2017/C3S_311a_Lot3_CNR;
GAIA-CLIM, the project funded from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 640276.

[1] Bodeker, G., et al., 2015: Reference upper-air observations for climate: From concept to reality. *Bull. Amer. Meteor. Soc.* 97, 123–135.

[2] Calculli C, Fassò A, Finazzi F, Pollice A, Turnone A. (2015) Maximum likelihood estimation of the multivariate hidden dynamic geostatistical model with application to air quality in Apulia, Italy. *Environmetrics.* 26(6), 406–417.

[3] Datta, A., Banerjee, S., Finley, A.O., and Gelfand, A.E. (2016), Hierarchical Nearest-Neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association.* 111, 800-812.

[4] Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, et al. 2001. The ERA-Interim reanalysis: configuration and performance of the data assimilation system, *Q. J. Roy. Meteor. Soc.* 137, 553–597.

[5] Delicado, P., Giraldo, R., Comas, C., & Mateu, J. 2010. Statistics for

spatial functional data: some recent contributions. *Environmetrics* 21(3-4), 224-239.

[6] Ignaccolo, R., Franco-Villoria, M., Fassò, A. 2015. Modelling collocation uncertainty of 3D atmospheric profiles. *Stochastic Environmental Research and Risk Assessment* 29 (2), 417-429.

[7] Fassò, A., Ignaccolo, R., Madonna, F., Demoz, B. and Franco-Villoria M. 2014. Statistical modelling of collocation uncertainty in atmospheric thermodynamic profiles. *Atmos. Meas. Tech.* 7, 1803–1816.

[8] Fassò, A., Negri, I., Finazzi, F., et al. 2017. Measurement mismatch studies and their impact on data comparisons. Deliverable D3.4, Technical report of GAIA project. http://www.gaia-clim.eu/page/deliverables.

[9] Finazzi, F.and Fassò, A. 2014. D-STEM: A Software for the Analysis and Mapping of Environmental Space-Time Variables. *Journal of Statistical Software* 62 (6), 1-29.

[10] Haimberger, L., Tavolato, C., and Sperka, S. 2012. Homogenization of the global radiosonde temperature dataset through combined comparison with reanalysis background series and neighboring stations, *J. Climate.* 25, 8108–8131.

[11] Kaufman, C.G., Schervish, M.J., Nychka, D.W. 2008. Covariance Tapering for Likelhood-Based Estimation in Large Spatial Data Sets, *Journal of the American Statistical Association.* 103, 1545-1555.

[12] Madonna, F., et al. 2017. Access to Baseline and Reference in-situ Observations, European Meteorology Society (EMS) annual meeting, EMS2017-846.

[13] Menafoglio, A., Secchi, P., Dalla Rosa, M. 2013. A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space. Electron. J. Statist. 7, 2209–2240.

[14] Merchant, C. J., Paul, F., Popp, T., Ablain, M., Bontemps, et al. 2017. Uncertainty information in climate data records from Earth observation, *Earth Syst. Sci. Data.* 9, 511-527.

[15] Porcu, E., Bevilacqua, M., Genton, M. G. 2016. Spatio-Temporal Covariance and Cross-Covariance Functions of the Great Circle Distance on a Sphere. *Journal of the American Statistical Association.* 11, 888–898.

[16] Ruiz-Medina, M.D. (2012) New challenges in spatial and spatiotemporal functional statistics for high-dimensional data. Spatial Statistics, 1, 8291

[17] Sherwood, S.C., Meyer C.L., Allen R.J., Titchner H.A. 2008. Robust tropospheric warming revealed by interactively homogenised radiosonde data. J. Clim. 21, 5336 – 5352.

[18] Stein, M.L. (2013), Statistical Properties of Covariance Tapers, *Journal of Computational and Graphical Statistics*, 22, 866-885.

[19] Thorne, P.W., Brohan, P., Titchner, H.A., et al. 2011. A quantification

of uncertainties in historical tropical tropospheric temperature trends from radiosondes. *J. Geophys. Res.* 116, D12116.

[20] Thorne, P. W., Madonna, F., Schulz, J., Oakley, T.,et al. 2017. Making better sense of the mosaic of environmental measurement networks: a system-of-systems approach and quantitative assessment, *Geosci. Instrum. Method. Data Syst. Discuss.* doi.org/10.5194/gi-2017-29, in press.

[21] Vecchia, A.V. (1988), Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society.* 50, 297-312.