Adaptive thinning of atmospheric observations in data assimilation with vector quantization and filtering methods – the first steps

Christoph Gebhardt[1], Tilo Ochotta[2], Dietmar Saupe[2], Werner Wergen[1]
[1]Deutscher Wetterdienst, [2]Universität Konstanz, Fachbereich Informatik und Informationswissenschaft
christoph.gebhardt@dwd.de

An important task in data assimilation for numerical weather prediction is the effective exploitation of large amounts of data produced by current and future observation systems, in particular satellite instruments. The high spatial and temporal density of these data is on the one hand highly valuable for estimating an initial state in the numerical forecast process. But on the other hand such data sets significantly increase the computational costs of the assimilation and, moreover, can violate the assumption of spatially independent observation errors and more complex observation error statistics would be needed leading to additional increase in the computational costs.
Thinning approaches which effectively reduce the number of analyzed observations but at the same time retain as much of the essential information content as possible can help to overcome theses problems. Liu & Rabier (2002) investigated assimilations of synthetic data with fixed spatial thinning intervals in a simplified model. They suggest that the optimal thinning which minimizes the analysis error depends on the spatial resolution of the model and of course on the degree of approximation of the observation error matrix with respect to its off-diagonal elements. The use of an 'influence matrix' of observations (Cardinali et al., 2003) is under investigation at ECMWF.
We develop thinning algorithms in an interdisciplinary project which are inspired by simplification methods from geometry processing in computer graphics and by clustering algorithms in vector quantization.
In a first method ('estimation error method'), we iteratively estimate the redundancy of the current data set and remove the most redundant observation. The degree of redundancy of an observation is defined to be inversely proportional to the interpolation error of its reconstruction obtained by applying an interpolation filter to a neighborhood in which the observation is removed.
In a second scheme ('cluster method'), the number of points in the output set is increased iteratively. These observations correspond to centers of clusters of observations. A distance measure that combines spatial distance with the difference in observation values defines an error measure for the overall quality of a clustering.
We apply the two methods to ATOVS satellite data which are processed in a 1D-Var scheme to retrieve profiles of atmospheric temperature and humidity. The profiles are again input data for the global analysis scheme of DWD.
Since the information gained by assimilating the data does not directly depend on the observations themselves but on their innovations to a first-guess state, we apply the thinning to the differences between bias-corrected observed brightness temperatures (TB_obs) and their first-guess (TB_FG, e.g. 3h-forecast temperatures transformed with RTTOV7) for those channels to be used in the 1D-Var scheme.
In a first test, we compared the two methods to a simple stepwise thinning which takes every third observation in zonal and meridional direction (*Figure 1*). For identical numbers of observations, the cluster method is spatially more homogeneous as the stepwise method which has the highest data density close to nadir view. The estimation error method leads to a more patchy thinning because is does not take into account the spatial distance as explicitly as the cluster method.
Figure 2 compares the three methods in terms of data density per $10.000 km^2$. The difference maps show that both cluster and estimation error method tend to increase the density compared to the stepwise thinning at the edges along the satellite track. This effect is more pronounced in the cluster method. In the region of overlapping tracks off the shore of southern Argentina with differing TB_obs-TB_FG due to time differences, the estimation error method results in the highest density of all methods. The thinning by the estimation error method is generally weaker in regions with more variable data because each observation in such an area is less likely to be redundant referring to its neighbors leading to small regions of comparably high data density. The cluster method also tends to retain more data in such regions but to a lesser degree due to its spatial constraint and the definition of a cluster center as some kind of balancing representative of its neighborhood. The histogram of data densities for 3387 observations confirms these effects. Compared to the stepwise thinning, the two other methods shift data density to higher values but significantly only up to 1 observation/10.000 km^2 for the cluster method but more effective for the estimation error approach. A density value for a distribution of 3378 points which is homogeneous over the area covered by the tracks can be approximated by 0.65 obs./10.000 km^2 which is close to the maximum of all methods in the histogram.
Tests with thinning down to lower numbers of retained observations reveal that the estimation error method tends to increase the variance of the innovation statistics. Consequently, thinning down to small data sets bears the danger of overemphasizing the extremes of TB_obs-TB_FG which on the one hand are possibly of most value in updating the first-guess but on the other hand are most affected by deficiencies of the preceding quality control of the data.
Although these first test results help to figure out the properties of the methods, the proper choice of method, degree of thinning, and parameters has to be determined in a complete assimilation and forecast experiment which is ongoing work.

Liu, Z.-Q. and Rabier, F. (2002): The interaction between model resolution, observation resolution and observation density in data ssimilation: A one-dimensional study. *Q.J.R. Meteorol. Soc*., **128**, pp.1267-1386.

Cardinali, C., Pezzulli, S. and Andersson, E. (2003): Influence matrix diagnostic of a data assimilation system. *ECMWF Seminar* 'Recent developments in data assimilation for atmosphere and ocean', 8-13 September 2003.

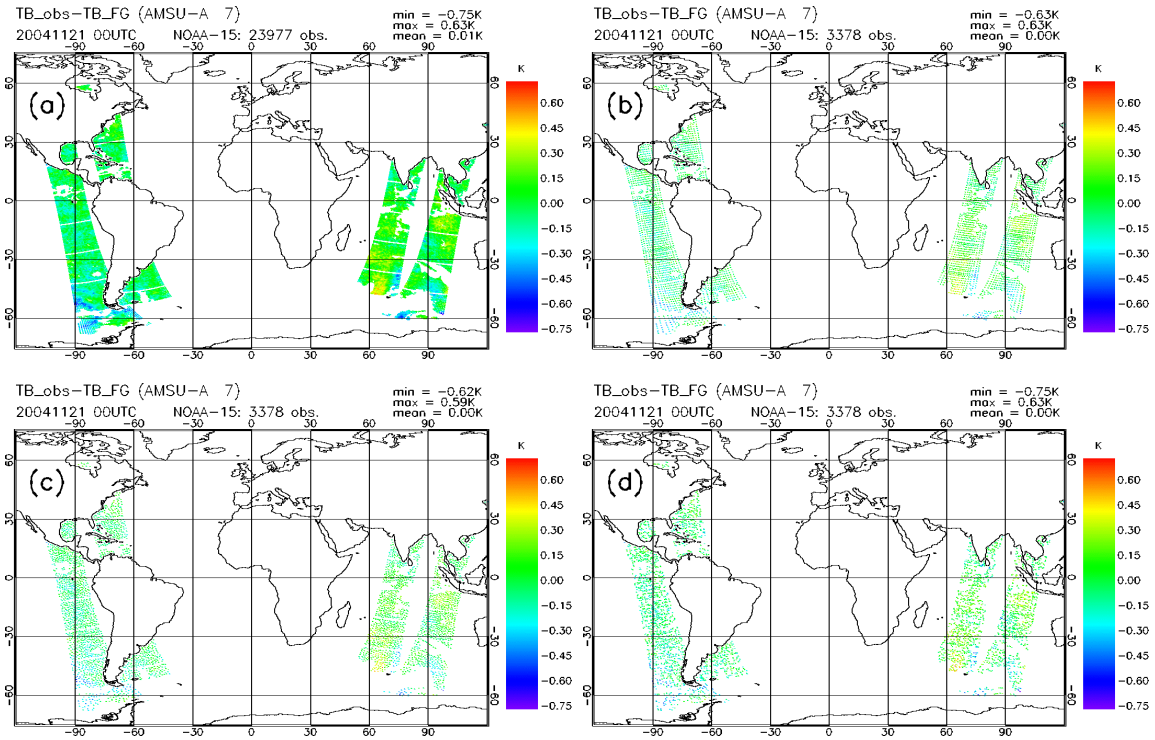

Figure 1: Bias-corrected minus first-guess brightness temperatures for channel AMSUA-A 7 of NOAA-15 on 21st Nov 2004 00 UTC for (a) unthinned data and after thinning by (b) stepwise, (c) cluster, and (d) estimation error method.

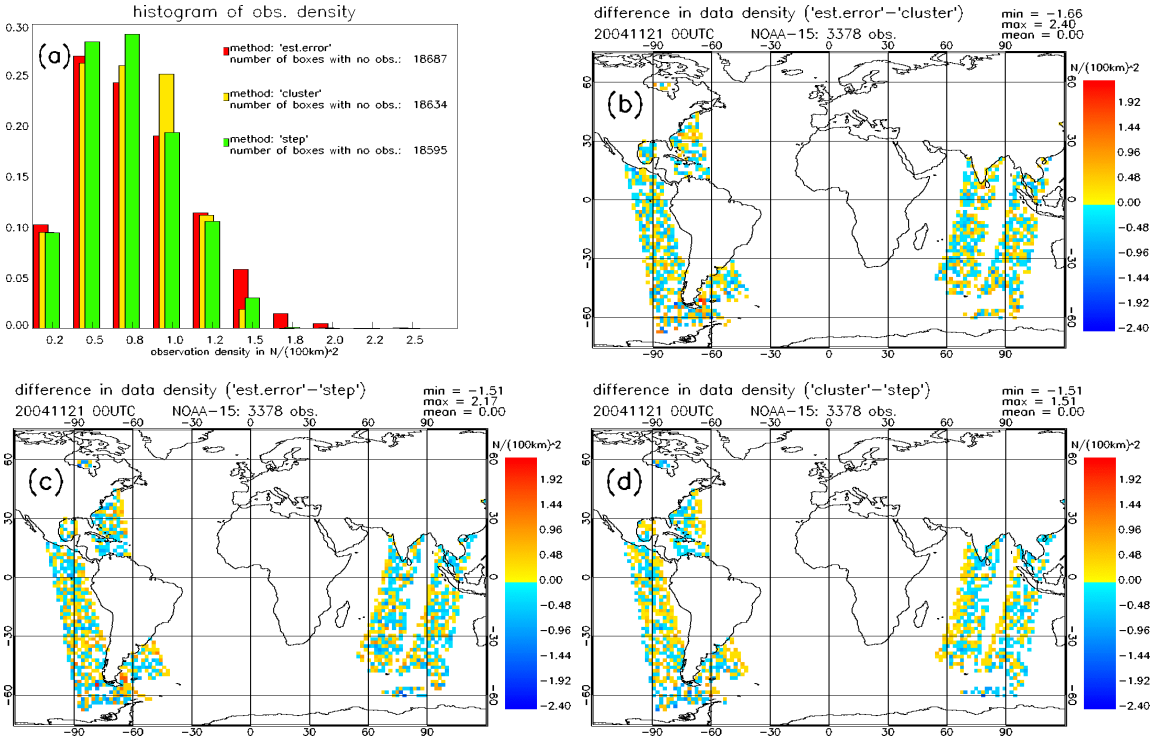

Figure 2: (a) histogram of data density after thinning with each method down to 3378 observations and difference maps of data density for (b) estimation error minus cluster, (c) estimation error minus stepwise, and (d) cluster minus stepwise method.