A different way to look at the intercomparison of datasets
illustrated with SCIAMACHY v5.02 versus lidar ozone profiles

J.A.E. van Gijsel, R. Zurita Milla, P. Stammes, S. Godin-Beekmann, T. Leblanc, M. Marchand, I.S. McDermid, K. Stebel, W. Steinbrecht and D.P.J. Swart

Anne.van.Gijsel@KNMI.nl & R.Zurita-Milla@UTwente.nl

In validation studies, the intercomparison of two datasets is traditionally carried out in predefined groups of observational characteristic like longitude, stellar magnitude or solar zenith angle. Here we present an alternative method in which we trained a self organizing map (SOM) with a full time series of relative difference profiles of SCIAMACHY and ozone profiles from seven NDACC lidars. Then, several SCIAMACHY and lidar data characteristics were mapped to the trained SOM. These maps were studied to see if the variation for a given characteristic explain the differences found between the datasets. Here we found altitude-dependent relations. From the lowest altitude studied (18 km) ascending, the most influencing factors were longitude, solar zenith angle, latitude and sensor age. At the highest altitudes (up to 45 km) these factors were the solar zenith angle and the day of the year. Clustering into three classes showed that there are also some local dependencies, with for instance one cluster having a much stronger correlation with the sensor age between 36 and 42 km. This novel SOM-based validation approach proved to be powerful and it is not being limited to a-priori defined data subsets.

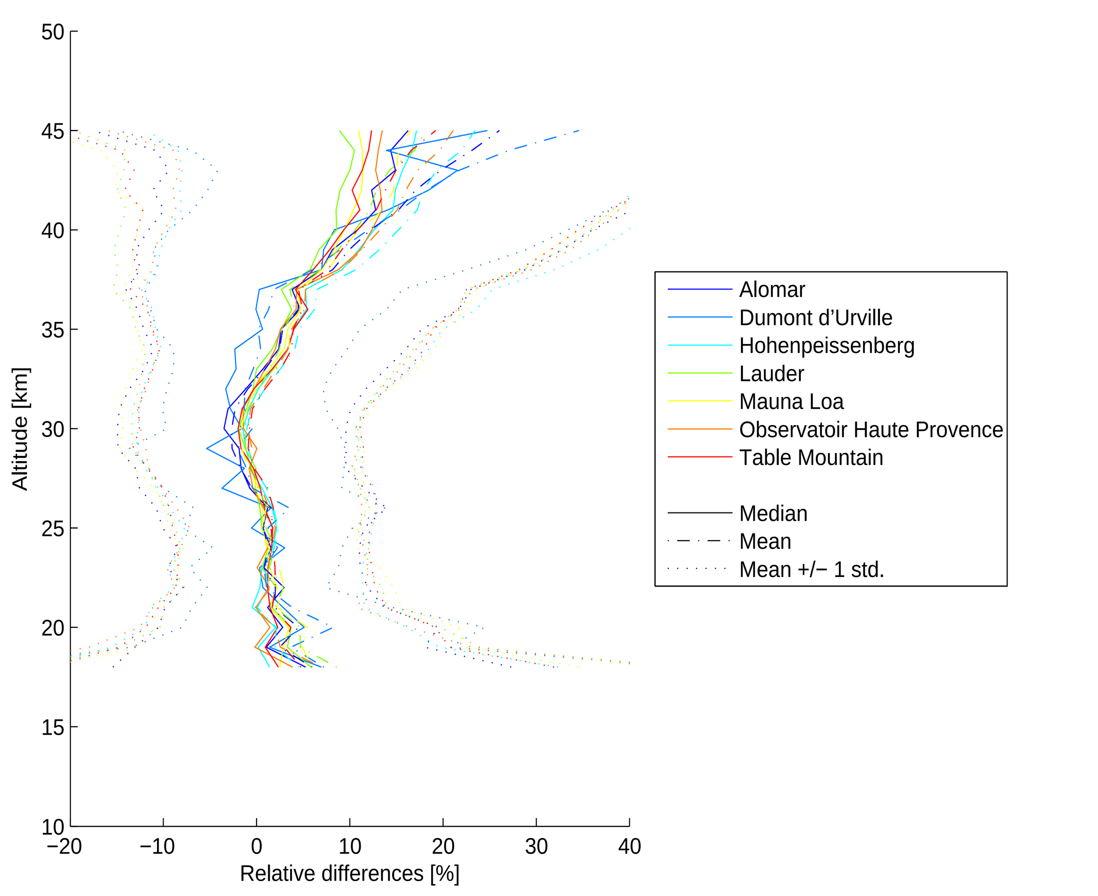

Figure 2. *Relative differences between SCIAMACHY v5.02 limb ozone profiles and lidar (location indicated by colour). The continuous lines correspond to the median differences, the dash-dotted lines to the mean differences and the dotted lines correspond to the mean ± one standard deviation.*

Step 1: Data gathering

Here we collocate lidar ozone profiles with SCIAMACHY v5.02 limb ozone profiles that are observed within 800 km and 20 hours. We obtain about 14000 collocated valid ozone profile pairs. Finally, the data are normalised before training. Fig. 2 shows the relative differences per lidar site of this data set.

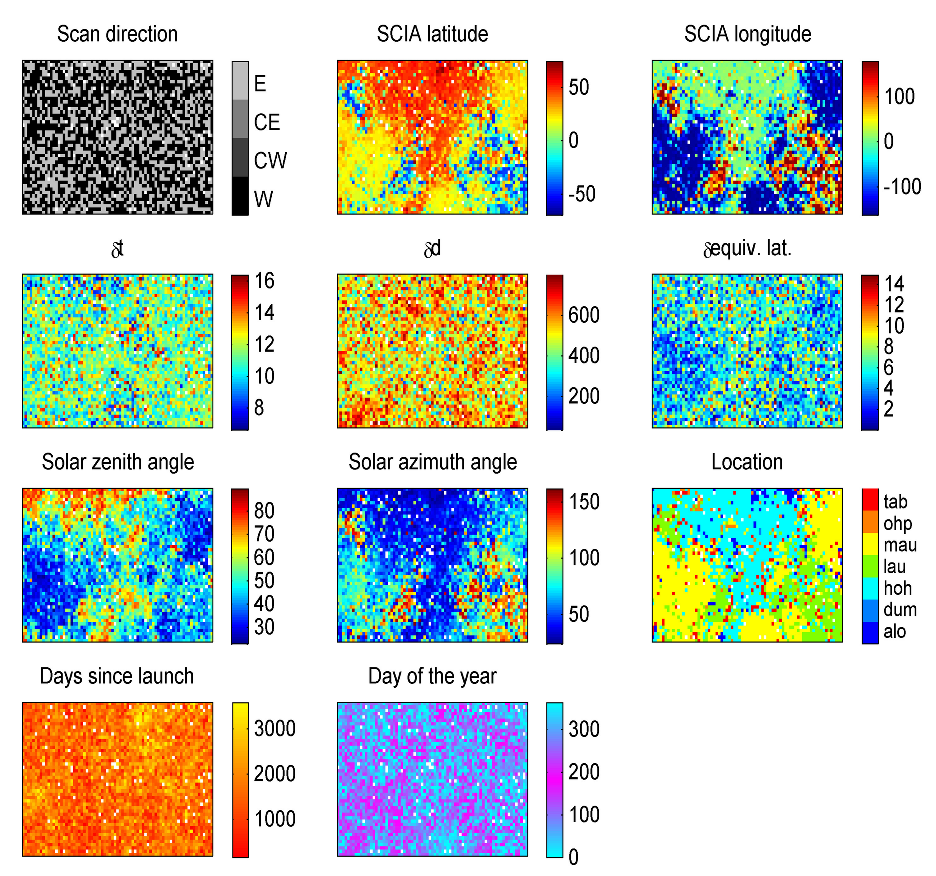

Figure 4. *EV planes showing f.l.t.r.t.t.b: SCIAMACHY scan direction, SCIAMACHY latitude, SCIAMACHY longitude, collocation differences in time, distance and equivalent latitude, SCIAMACHY SZA and SAA, lidar site, Number of days since the start of the mission and DOY*

Step 3: Mapping EVs onto the SOM

Fig. 4 shows the mean value of each EV for all input vectors that map to a specific neuron. Some EVs show clear patterns, other show local clusters and some do not exhibit any structure. Some of the patterns show linking of multiple EVs.

www.atmos-meas-tech.net/8/1951/2015

Methodology

The flowchart below describes the steps from data selection and preprocessing, training of the SOM, mapping of the codebook vectors and explanatory variables (EVs), clustering to correlation analysis

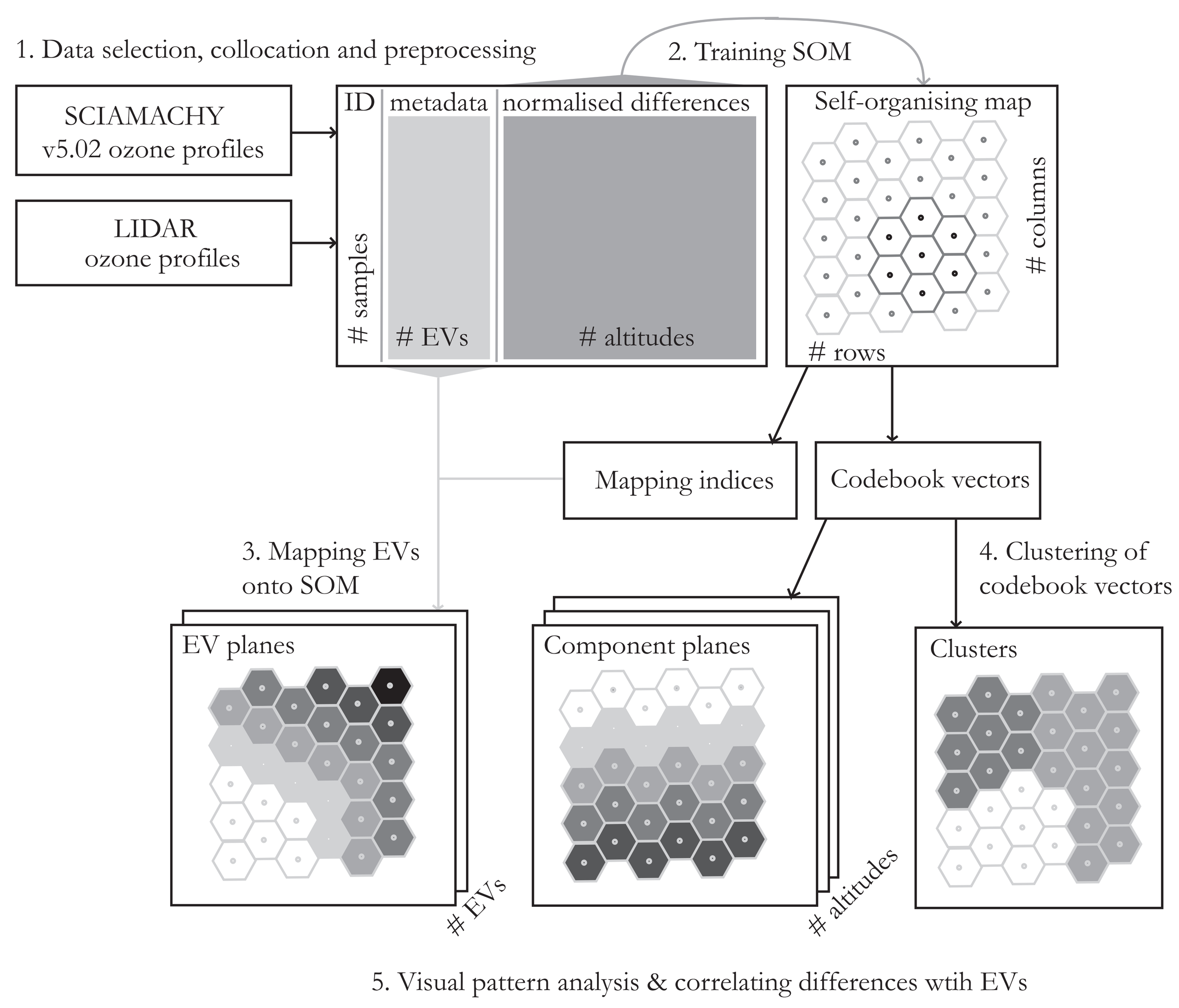

Figure 1. *Flowchart of the implementation steps:*
1. *Data selection, collocation and preprocessing*
2. *Training of the SOM using the profile differences*
3. *Mapping the EVs onto the SOM*
4. *Clustering of the codebook vectors*
5. *Visual pattern analysis and correlating the organised differences with the EVs*

Steps 3 and 4 can also be done in reversed order.

Step 5: Pattern & correlation analysis

The correlation R is computed between the codebook vectors and the mapped EVs. Fig. 6 shows the R for each altitude bin and EV and Fig. 7 does the same for the three identified clusters. See abstract above and the full paper for a discussion of these findings ☺.

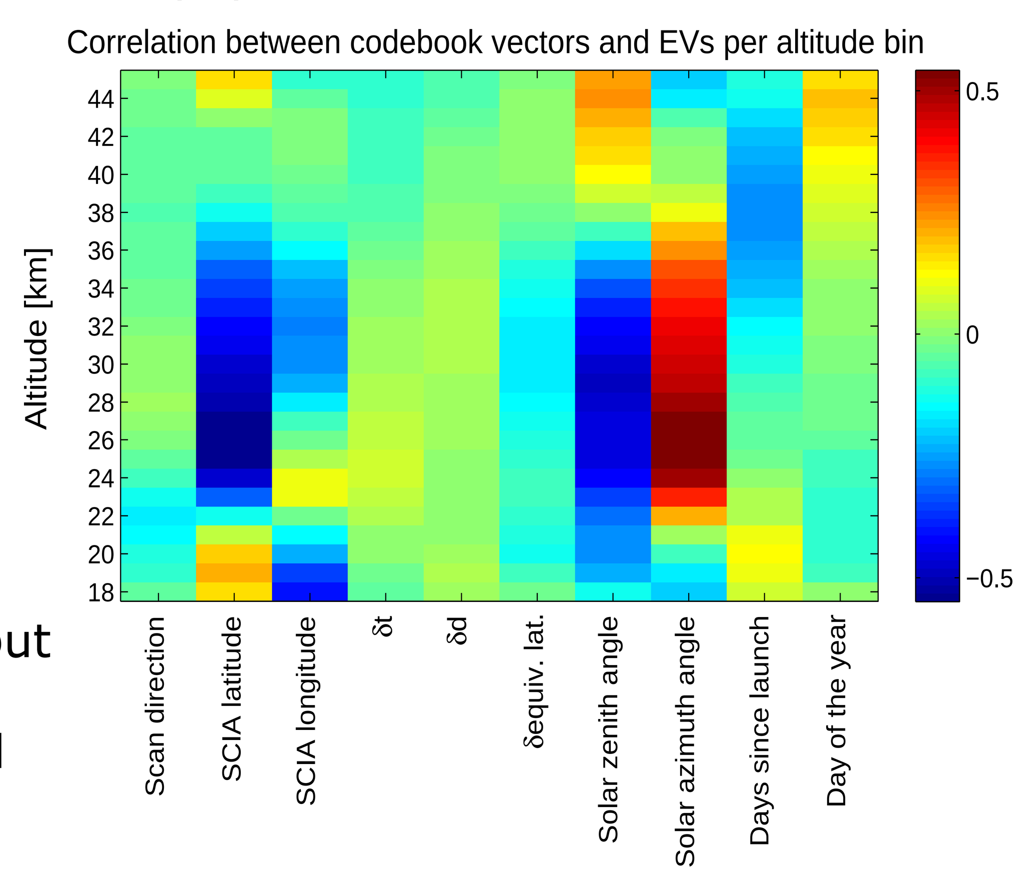

Figure 6. *Correlation between the organised differences and the mapped EVs of the corresponding data*

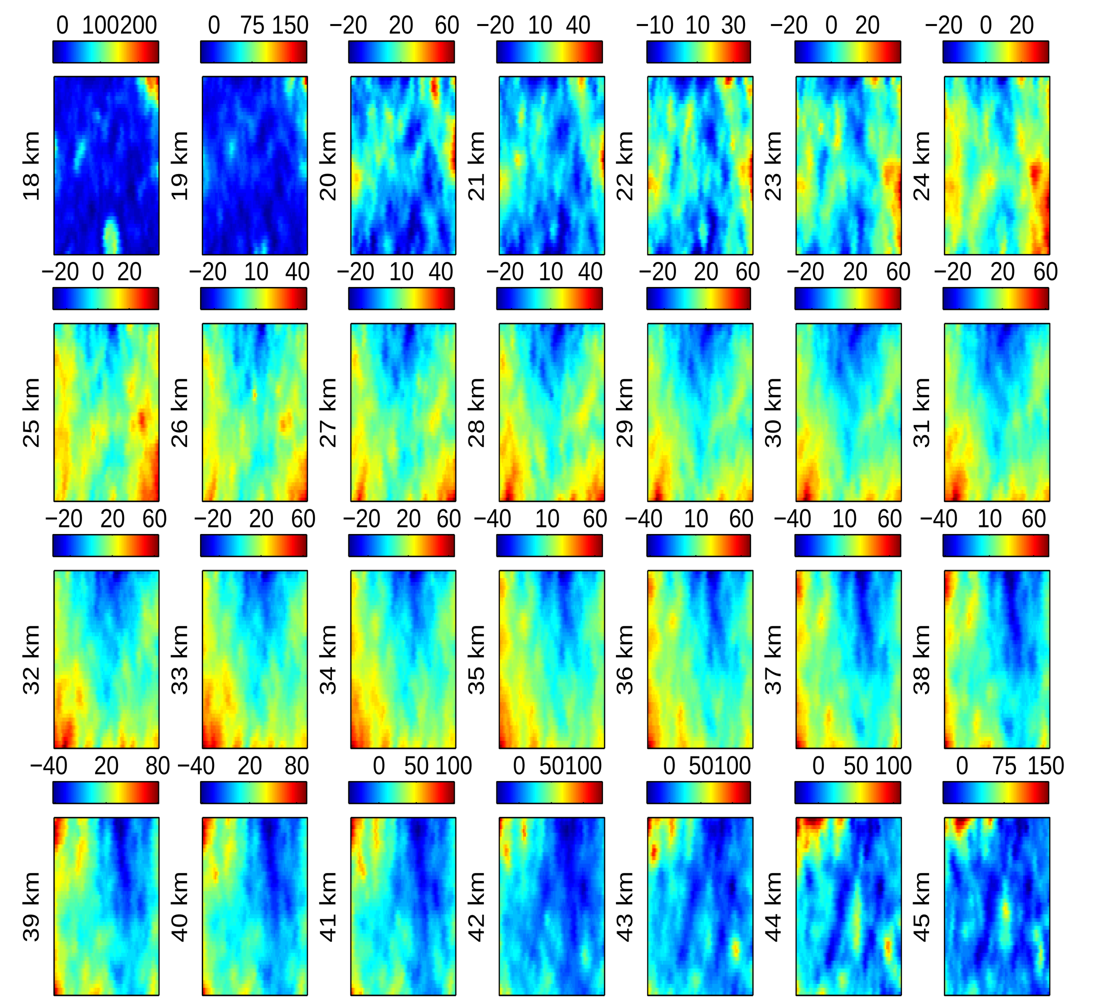

Figure 3. *Component planes with the mapped ozone differences denormalised to the original ranges. Altitudes are organised from 18 km (top left) to 45 km (bottom right).*

Step 2: Training of the SOM

Here we chose to create a lattice grid of 46 × 75 neurons. Training was done in two phases. Fig. 3 shows how the relative differences in ozone concentrations were mapped for each 1-km altitude bin.

Step 4: Clustering of the codebook vectors

Consulting various indices, the optimal number of clusters were determined to be three. Fig. 5 shows the resulting clusters after k-means classification of the codebook vectors in Fig. 3.

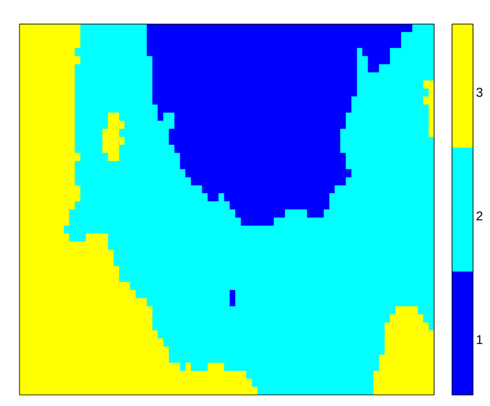

Figure 5. *Locations of the clusters obtained by k-means classification of the codebook vectors.*

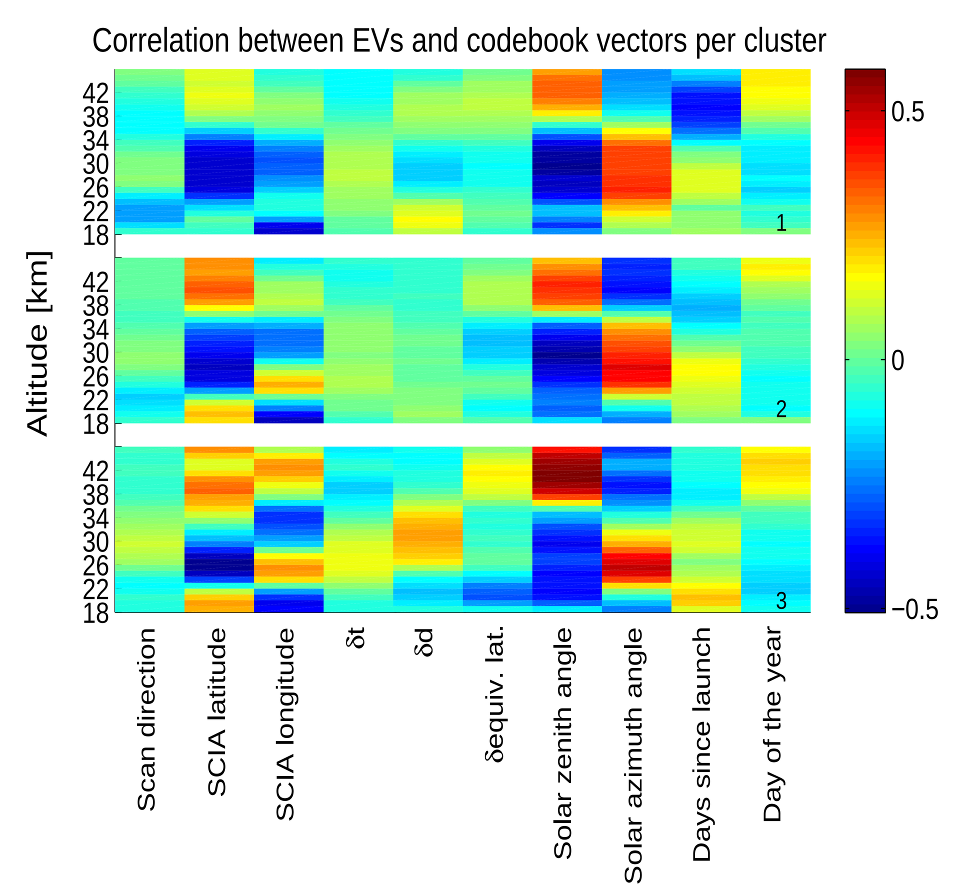

Figure 7. *Correlation between the organised differences and the mapped EVs of the corresponding data for each of the three clusters in Fig. 5.*