Supplement of Atmos. Chem. Phys., 15, 6943–6958, 2015
http://www.atmos-chem-phys.net/15/6943/2015/
doi:10.5194/acp-15-6943-2015-supplement

Atmospheric Chemistry and Physics Open Access

Supplement of

On the competition among aerosol number, size and composition in predicting CCN variability: a multi-annual field study in an urbanized desert

E. Crosbie et al.

Correspondence to: A. Sorooshian (armin@email.arizona.edu)

Explanation of size distribution shape clustering method

Each size distribution observation comprises 112 number concentrations, n, corresponding to the discrete bins reported by the SMPS software on a logarithmic, regularly spaced grid. Each observation is considered as a vector, $\mathbf{x}_i$, normalized using the root sum of squared distance (Euclidean norm) to isolate the effect of size distribution shape from the (strong) variability in total CN:

$$\boldsymbol{x}_i^* = \{n_1, n_2, \dots, n_{112}\}$$

$$\boldsymbol{x}_i = \boldsymbol{x}_i^* / \|\boldsymbol{x}_i^*\|_2$$

The K-means algorithm seeks the optimal position of a predetermined number, k, of centroid vectors, $\mathbf{c}_j$ which minimizes the residual distance of observation vectors assigned to that cluster set, C_j, from the centroid. Centroids are initially randomly assigned to observations and then temporary cluster assignments are made to the closest centroid for each observation, $\mathbf{x}_i$. The centroid is then recalculated based on the cluster members and the process is iterated until convergence. The method is known to sometimes converge on local optima rather than global optima in some cases and so the operation is repeated many times (>100) with different random starting observations to check for global convergence.

$$C_j^{(m)} = \left\{ \boldsymbol{x}_i : \left\| \boldsymbol{x}_i - \boldsymbol{c}_j^{(m)} \right\|_2^{\,2} \le \left\| \boldsymbol{x}_i - \boldsymbol{c}_p^{(m)} \right\|_2^{\,2} \quad \forall p \right\}$$

$$\boldsymbol{c}_i^{(m+1)} = \frac{1}{\left| C_j^{(m)} \right|} \sum_{\boldsymbol{x_i} \in C_j^{(m)}} \boldsymbol{x_i}$$

The resulting cluster centroids after convergence represent the mean size distribution shape of the subset of observations attributed to that cluster. Observations show a quasi-continuous nearest neighbor density in the regions between the centroids, hence using the "hard" cluster associations (as per the traditional K-means algorithm) creates an abrupt transition for observations which lie close to the cluster boundary. Instead the cluster association is made "fuzzy" using cluster assignment weights, w_j. Weights are based on inverse squared Euclidean distance to each centroid, c_j, except where the square distance is greater than the sum of squared distances via any other centroid, c_p, in which case the weight, w_j, is zero. Weights are normalized such that $w_j = 1$.

$$d_{i,j} = \left\| \boldsymbol{x}_i - \boldsymbol{c}_j \right\|_2$$

$$w_{i,j}^* = \begin{cases} d_{i,j}^{\,-2}; & d_{i,j}^{\,2} \le d_{i,p}^{\,2} + \left\| \boldsymbol{c}_p - \boldsymbol{c}_j \right\|_2^{\,2} \\ 0; & d_{i,j}^{\,2} > d_{i,p}^{\,2} + \left\| \boldsymbol{c}_p - \boldsymbol{c}_j \right\|_2^{\,2} \end{cases}$$

$$w_{i,j} = w_{i,j}^* \Big/ \sum_j w_{i,j}^*$$

The chosen number of cluster centroids (4) was selected subjectively to balance the improvement in model performance by increased numbers of clusters with the simplicity of retaining as few as possible. The quality of the K-means model as the number of centroids is changed can be visualized by the sum of residual distances between observations and closest cluster centroid. This is shown in Figure S1. The cluster centroids were also identified using only the data from the summer seasons (PM and M) since these seasons contain the largest range of variability in the size distribution shape. Repeating the K-means algorithm using the entire dataset finds the clusters common in winter (FF and WN) but finds a hybrid of the more extreme summertime clusters (N and CC) with the FF and WN clusters, respectively. Since the goal is to be able to isolate features, the clusters based on the summer "training" data are retained. Associations based on the fuzzy logic described above are then assigned to all data.

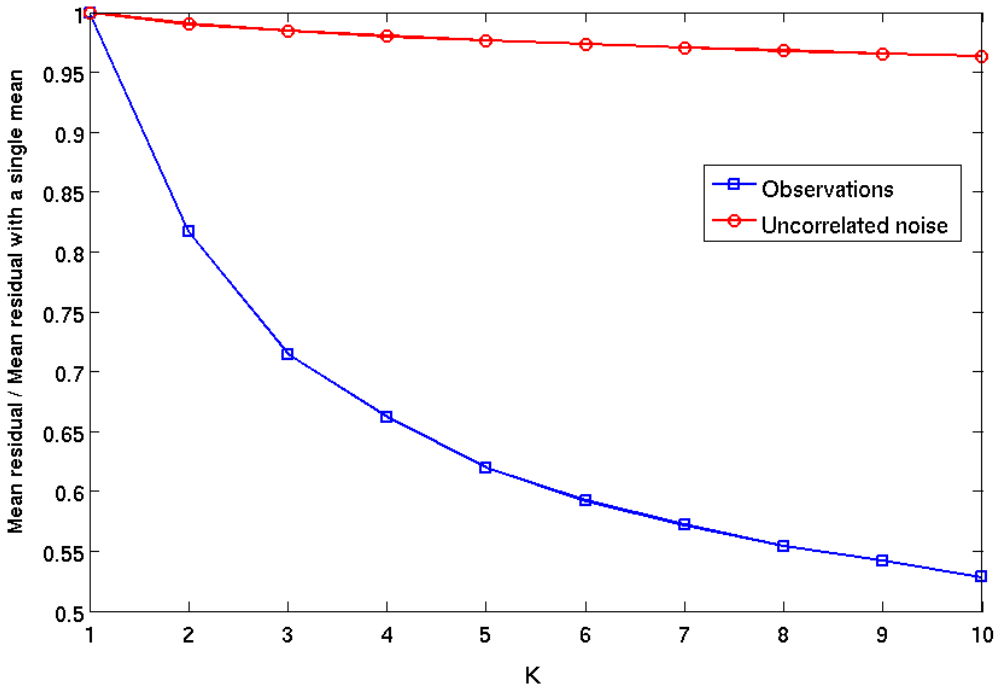

Figure S1: Performance of the K-means method as a function of the number of centroids, k. The residual is measured as the Euclidean norm of the vector difference between the observation and its allocated centroid and the mean is generated over all "training" observations. For comparison, the performance of the K-means model is shown for uncorrelated random noise and highlights the latent structure of the observations.