Cluster Analysis: A New Approach Applied to Lidar Measurements for Atmospheric Boundary Layer Height Estimation

DANIEL TOLEDO, CARMEN CÓRDOBA-JABONERO, AND MANUEL GIL-OJEDA

Atmospheric Research and Instrumentation Branch, Instituto Nacional de Técnica Aeroespacial, Torrejón de Ardoz, Madrid, Spain

(Manuscript received 10 December 2012, in final form 1 November 2013)

ABSTRACT

Several procedures are widely applied to estimate the atmospheric boundary layer (ABL) top height by using aerosols as tracers from lidar measurements. These methods represent different mathematical approaches, relying on either the abrupt step of the aerosol concentration between the ABL and the free troposphere (FT) or the statistical analysis of vertical variations of the aerosol concentration. An alternative method—the cluster analysis (CA)—has been applied to lidar measurements for the first time, emerging as a useful and robust approach for calculating the ABL height, taking the advantage of both previous variables: the vertical aerosol distribution as obtained from the lidar range-corrected signal (RCS) and the statistical analysis of the RCS profiles in terms of its variance to determine a region of high aerosol loading variability. CA limitations under real situations are also tested, and the effects in ABL height determination of both noise and cloud contamination in RCS are examined. In particular, CA results are weakly sensitive to the signal noise due to the basic features of this statistical method. In addition, differences in the ABL top height, as estimated under cloudy and clear skies, have been found to be lower than 1.8% for a high RCS signal, while no effect is observed for low RCS cloud conditions. Moreover, the CA performance on the ABL top height determination for real cases is also presented, showing the reliable CA skills in reproducing the ABL evolution.

1. Introduction

The atmospheric boundary layer (ABL) is the lowest part of the troposphere that is directly influenced by the earth’s surface with a time scale of 1 h or less (Stull 1988). Factors such as the season, orography, time of the day, and weather act over the ABL. A good knowledge of the determination of the ABL height is required for pollution dynamics studies and weather forecasting modeling (Seibert et al. 1997, 2000). Radiosoundings, typically used for ABL estimation and based on water vapor mixing ratio (Heffter 1980; Stull 1988), temperature, and wind and pressure profiles, are launched on a four-per-day basis, at best. Therefore, knowledge of the temporal evolution of the ABL during the day is poorly provided.

Active remote sensing systems such as lidars use aerosols as tracers to estimate the ABL top height (Russell et al. 1974; Lammert and Boesenberg 2006; Morille et al. 2007) because higher concentrations of aerosols are usually present in the ABL with respect to those in the free troposphere (FT). Indeed, a sharp discontinuity in lidar backscattered signal is usually found by “crossing” from ABL to FT altitudes. In addition, ABL heights are easily reached within the measurement range of most lidar systems. Moreover, since lidars can be in continuous operation with a vertical resolution of a few meters and an integration time of seconds to minutes, they appear to be the most appropriate instrumentation for continuous ABL top monitoring.

Several lidar approaches for ABL height detection have been developed. Among them, Melfi et al. (1985) and Boers et al. (1988) defined the ABL top as the height point where the backscattered signal drops below a given threshold; however, the drawback of this method is indeed the selection of an appropriate threshold value. Other authors (Hooper and Eloranta 1986; Piironen and Eloranta 1995; Menut et al. 1999) used the height where the first maximum of the signal variance is found to establish the ABL top height. In addition, derivative
methods, more commonly used for that purpose, analyze the derivative of the lidar signal as a function of height: 1) the simple derivative (Hayden et al. 1997; Flamant et al. 1997); 2) the derivative of its logarithm with respect to height (Senff et al. 1996); and 3) the second derivative of the signal with respect to height (Menut et al. 1999). These methods, assuming also a higher aerosol concentration in ABL than in FT, identify the ABL top by the height point corresponding to the absolute minimum of the first or second derivatives of the lidar signal. A comparison analysis of all these methods is found in Sicard et al. (2006). However, these derivative methods suffer from the noise effect, and small-scale structures in the lidar signal introduce large gradients unrelated to the ABL top height.

Another approach, wavelet covariance transform (WCT) (Steyn et al. 1999; Cohn and Angevine 2000; Davis et al. 2000; Hägeli et al. 2000; Brooks 2003), which is less affected by the signal noise, is also used for ABL top determination. WCT, also called Haar wavelet, provides a multiscale local gradient analysis of the lidar signal to isolate the spikes due to aerosol concentrations. WCT is based on the Haar function that depends on two parameters: translation and dilation (i.e., Brooks 2003). It is well known that the ABL top height is located where the maximal WCT value is found for appropriate values of those two parameters dependent on the depth of the transition zone. A recently published comparative study (Pal et al. 2010) shows the ABL height as obtained by the Haar wavelet method lies between the height estimated by the logarithm of the derivative of the signal and its second derivative, both with respect to height.

In summary, all of these methods are based on two approaches: 1) the vertical distribution of the aerosol concentration to find a significant transition, as applied by the threshold method, derivative methods, and the WCT; and 2) the statistical analysis as used by the maximum of the signal variance.

A new procedure to determine the ABL top height by using lidar measurements combining both of these approaches is presented in this work: the cluster analysis (CA). CA has been widely applied to both meteorology and climatology studies, as reported in different works (Fernau and Samson 1990; Cheng and Wallace 1993; Fovell and Fovell 1993; Bunkers et al. 1996; De Gaetano 1996). However, this statistical method is, to our knowledge, first applied to lidar measurements for the ABL top height estimation in this work.

The goal of this work is the development of a useful and robust method for calculating the ABL top height by using both the vertical distribution of aerosols and the variation of their concentration. In this sense, the CA method is described in detail in section 2. The main results and discussion are exposed in section 3: the application of the CA to lidar measurements in order to calculate the ABL top height is described in section 3a, and the effects of the lidar signal noise and the presence of cloud contamination in the lidar backscattered signal in the estimation of the ABL height by using CA are exposed in sections 3b and 3c, respectively. In addition, the CA performance on the ABL top height determination for real cases is exposed in section 3d, with a brief description of the lidar system used for aerosol measurements. Finally, the main conclusions are summarized in section 4.

2. Cluster analysis: Description of the method

The CA method is mainly based on assembling a set of objects into groups (Anderberg 1973), where the objects in the same cluster are similar and objects in different clusters are dissimilar (Balling 1984). In particular, the distance between objects is the similarity used to form the clusters. These distances (similarities) are based on single or multiple dimensions, where each dimension represents a rule or condition in order to group the objects. CA can be hierarchical and non-hierarchical depending on the way the clusters are formed.

a. Hierarchical cluster analysis

Hierarchical clustering is a method based on building a hierarchy of clusters, by means of two strategies: divisive and agglomerative (Theodoridis and Koutroubas 1999). Divisive hierarchical CA begins with all data points initially grouped into a single cluster and splits are performed recursively as one moves down the hierarchy, until finally each point represents its individual cluster (Johnson and Wichern 1992). Agglomerative hierarchical CA starts with each observation in its individual cluster, and pairs of clusters are merged as one moves up the hierarchy, until the final step when all data points remain in the same cluster (Johnson and Wichern 1992). The agglomerative hierarchical cluster analysis algorithm process to form the clusters is shown in Table 1. To classify the observations in different clusters, both hierarchical strategies (as discussed below) require a criterion to stop the process—that is, each strategy finishes the process when either only one cluster remains or all the observations represent a unique cluster, indicating no classification.

Therefore, the distance between two points, i and j, and the distance between clusters must be defined in both hierarchical and nonhierarchical CA methods. The Euclidean distance is the most frequently used
The hierarchical clustering method is illustrated in the left panel of Fig. 1.

As mentioned before, a criterion must be selected to stop the hierarchical CA process, since the number of clusters is not required to be a priori specified. For this purpose, hierarchical CA is graphically displayed by using a tree diagram, called dendrogram, representing that process. A dendrogram scheme for a dataset of four observations is shown in Fig. 2. Information about the compactness, the distinctness, and the number of observations for each cluster is obtained by examining the dendrogram. For instance, the number of observations for each cluster is given by the number of leaves in the branch of the dendrogram. Vertical and horizontal axes in Fig. 2 represent the distance and the distinctness, respectively, between clusters. Therefore, compact and distinct clusters are presented at the scheme with large distances in both the vertical and horizontal axes.

The total time required for a hierarchical clustering process is of \( O(N^2 \log N) \), where \( N \) is the total number of points. Then, this time complexity limits the amount of data to be processed.

b. Nonhierarchical cluster analysis (K-means method)

In nonhierarchical CA methods, \( K \) means is the most used algorithm (MacQueen 1967). The number of clusters is a priori denoted and points, called cluster seeds, are random or follow some heuristic inserted into the dataset. Therefore, the number of cluster seeds corresponds to the number of clusters selected.

Each cluster seed acts as an initial cluster center. Clusters are formed by assigning each observation to the nearest seed. After all observations are assigned, the cluster seeds are replaced by the cluster center. This process (see Fig. 1, right) is repeated until the changes in the cluster center become closer to zero. In this work, Euclidean distance is considered as a measure of similarity between sets of observations. The cluster center is identified as the point that minimizes the sum of the squared errors (SSE) function, defined as

\[
SSE = \sum_{i=1}^{K} \sum_{x \in c_i} \text{dis}(c_i, x)^2
\]

where \( K \) is the number of clusters, \( \text{dis} \) is the Euclidean distance, \( c_i \) is the cluster \( i \), and \( x \) is an observation in the cluster \( i \). Therefore, the cluster center is the mean distance between observations in the cluster. The \( K \)-means algorithm process is shown in Table 2.

The final results obtained by this method depend on the number of initial cluster centroids selected. Thus,
previous knowledge of the number of clusters present in the dataset is required for cluster validation, since the number of clusters is an a priori parameter to be introduced into the cluster analysis algorithm. Therefore, results must be validated in terms of quantities involving the vectors of the dataset themselves. All the criteria in cluster validation have a common objective in order to find the clustering, where the observations are in compact and well-separated clusters.

For this purpose, two indexes widely applied in cluster validation analysis are used in this study: Dunn (Dunn 1974) and Davies–Bouldin (Davies and Bouldin 1979) indices. The Dunn index $D$ is defined as

$$D = \min_{1 \leq i \leq n} \left( \min_{1 \leq j \leq n, j \neq i} \left\{ \frac{d(c_i, c_j)}{\max_{1 \leq k \leq n} [d^*(c_k)]} \right\} \right), \quad (5)$$

where $d(c_i, c_j)$ is the distance between clusters $c_i$ and $c_j$, representing a measure of dissimilarity; $d^*(c_k)$ is the intracluster distance of cluster $c_k$, as a measure of dispersion of the cluster; and $n$ is the number of clusters. In case observations contain compact and well-separated clusters, large values are found for $d(c_i, c_j)$, and $d^*(c_k)$ is expected to be small. Therefore, the presence of compact and well-separated clusters is indicated by large $D$ values.
On the other hand, the Davies–Bouldin index DB is defined as

$$DB = \frac{1}{n} \sum_{i=1}^{n} \max_{i \neq j} \left( \frac{S_n(Q_i) + S_n(Q_j)}{S(Q_i, Q_j)} \right),$$

where $S_n$ is the averaged distance between all the observations of the cluster and the center of the cluster $Q$, $S(Q_i, Q_j)$ is the distance between cluster centers $Q_i$ and $Q_j$, and $n$ is the number of clusters. In this work, the cluster center $Q$ is defined as the average of the observations belonging to the same cluster. Therefore, the DB index is a measure of similarity between each cluster and its most similar one, and the number of clusters that minimizes the DB index is considered the optimal one.

In summary, an optimal number of clusters is a priori estimated for validation cluster analysis by using both $D$ and DB indexes. The total time required for $K$-means processing is linear in respect to the number of data points. In particular, this time is $O(I K n s)$, where $I$ is the number of iterations required for convergence, $K$ is the number of clusters, $n$ is the number the attributes, and $s$ is the number of observations. Therefore, nonhierarchical CA is more adequate for processing large datasets, whereas hierarchical CA is more suitable for moderate ones.

c. Data normalization

In the case that the magnitude of the variables is quite different, such variables must be normalized. Data normalization is based on forming z scores $Y_i$ from the given variables $X_i$ following the expression

$$Y_i = \frac{X_i - \bar{X}}{\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}},$$

where $\bar{X}$ is the mean value of those variables $X_i$, and $\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}$ is their standard deviations, where $n$ is the number of observations in the dataset. Another expression also used for data normalization is

$$Y_i = \frac{X_i}{\max(X_i)},$$

where $\max(X_i)$ is the maximum of all the $X_i$ values.

Therefore, data normalization provides that all variables contribute equally by forming the different clusters. In particular, Eq. (7) is used in this work.

3. Results and discussion

a. ABL height estimation by CA as applied to lidar measurements

The entrainment zone (EZ) is an interfacial layer between the ABL and the FT, where clear air masses from the FT are mixed with those more turbid coming from the ABL. Thus, the variability of the aerosol concentration in that EZ is very high. On the basis of this fact, the variance method (VM) uses a statistical analysis in order to analyze the variability of the aerosol concentration. In particular, the averaged ABL “centroid” height (Hooper and Eloranta 1986) is identified as the lowest height position of a local maximum in the variance function $VAR(Z)$ profile, defined as

$$VAR(Z) = \frac{1}{(N-1)} \sum_{i=1}^{N} [S(Z, t_i) - \bar{S}(Z)]^2,$$

where $S(Z, t_i)$ is the range-corrected lidar backscattered signal (RCS) at time $t_i$ and height $Z$, and the upper bar means the averaged profile from $N$ RCS profiles. The height position corresponding to the absolute maximum of the $VAR(Z)$ function establishes the center of the transition zone. Moreover, a vertical gradient in RCS is found at the transition zones due to the fact that higher aerosol concentrations are found in the ABL. By assuming that the EZ corresponds closely with the temperature inversion (Brooks 2003), the top of the ABL is identified as the center of the EZ layer (Stull 1988).
For instance, an idealized lidar profile is examined corresponding to a height-independent RCS in the ABL, followed by a smooth decrease of RCS in the entrainment zone, and a slightly height-independent RCS in the FT. That idealized lidar RCS profile, obtained from a given set of RCS profiles as averaged in time, and its variance are illustrated in Figs. 3a and 3b, respectively. Methods based on the vertical distribution of aerosols are focused on finding a transition zone where the RCS gradient shows a clear minimum and the VAR(Z) function shows a clear maximum. This height position corresponds to the ABL top height. Consequently, both the vertical aerosol distribution and its variance are needed in order to estimate the ABL height. However, at present, there is not a suitable method to calculate the ABL height by using both these quantities. A mean profile and its variance are estimated by using $N$ profiles, and each observation is represented in three dimensions, where the height, lidar RCS, and variance of the signal are the variables. As the variance shows a clear maximum at the ABL top height and the ABL contains a higher aerosol concentration than the FT, CA is used to estimate the ABL height as a function of both the vertical distribution and the variation of their concentration. To form clusters, the distance between two observations must be defined. In this study, Euclidean distances are used, and for simplicity, the distance $\text{dis}(i, j)$ between two given signal points, or observations, $i$ and $j$, with coordinates ($\text{height}_i$, $\text{RCS}_i$, $\text{VAR}_i$) and ($\text{height}_j$, $\text{RCS}_j$, $\text{VAR}_j$), respectively, is defined as

$$\text{dis}(i, j) = [(\text{height}_i - \text{height}_j)^2 + (\text{RCS}_i - \text{RCS}_j)^2 + (\text{VAR}_i - \text{VAR}_j)^2]^{1/2}. \quad (10)$$

Different clusters are formed by using these distances, and normalized values of all the variables (see section 2c) contribute equally to form them. Therefore, the distribution of the aerosols in the atmosphere and variations of their concentration can be studied by this method.

In that idealized lidar profile and its variance as illustrated in Figs. 3a and 3b, respectively, the scale is arbitrary, and all the variables were normalized (see section 2c). The gradient of the lidar RCS takes a negative minimum at the ABL top height. Moreover, the variance of the lidar RCS profile is also assumed to show a clear maximum at the ABL top height. The $K$-means approach is applied to the idealized lidar RCS profile, where two cluster seeds are selected to form the clusters in the basis of both the $D$ and DB validity indexes (see section 2b). Similar results were obtained by using hierarchical cluster analysis (see section 2a).

Results obtained by using the CA method for the case shown in Figs. 3a and 3b are presented in Fig. 3c, where two clusters are differentiated: points with the same color belong to the same cluster (as shown by black and white points, respectively, in Fig. 3c). In this sense, clusters represent the division of the atmosphere in the two zones depending on both aerosol distribution and concentration variability. The ABL top is localized at the height point where a change from a cluster to another one

![Fig. 3](https://example.com/fig3.png)

**Fig. 3.** (a) Idealized profile of a lidar RCS with only a vertical gradient at the ABL top. (b) The corresponding profile of the variance computed from successively measured profiles; the variance shows a clear maximum at the ABL top height. (c) Results obtained by using the CA method, where two clusters are differentiated, as shown by black and white points. The ABL is shown with a gray-shaded band respect to the FT.
is observed. The selection of an appropriate number of clusters is the main challenge for a successful result of the ABL height estimation by using CA. Therefore, cluster validity is an important procedure to obtain information about the ABL height.

The discussion exposed above relates to highly idealized smooth profiles, where the minimum and maximum of the first (or second) derivative of the lidar RCS and its variance, respectively, are located in the center of the transition zone. However, for real lidar RCS profiles, small-scale structure into the ABL could lead to differences in height between different methods.

The ABL top heights obtained under certain conditions by applying the VM are higher than those obtained by methods considering only the vertical aerosol distribution. This is caused by the role played by the entrainment process. Methods based on the vertical distribution of the aerosol concentration obtain the height point where the particle concentration decreases, whereas particle scattering is the most variable for methods based on the statistical analysis. Therefore, VM overestimates the ABL top height, since the height point of the local standard deviation maximum appears above the larger concentration decrease observed. The thickness of the entrainment layer is related to the difference in height between VM and methods based on the vertical aerosol distribution. In these cases, the ABL top height is difficult to be accurately established.

A second idealized lidar RCS profile and its variance are illustrated in Figs. 4a and 4b, respectively. In this case, the height calculated by only considering the vertical distribution of the aerosol concentration differs from that obtained by VM. Results obtained by using the CA method (see Fig. 4c) indicate that the ABL height point obtained by CA is located between those heights obtained by the other two methods. Therefore, CA, by using both the vertical aerosol distribution and its variance to obtain the ABL top height, makes available a single result when other methods provide different ones.

b. The effect of signal noise

The increase of noise with height influences the estimation of the ABL height. Since gradient and variance methods estimate the ABL height studying the vertical signal gradients and the vertical variance of the signal along the time, they are strongly affected by the signal noise. Sometimes it is necessary to smooth the lidar signal before estimating the ABL height. CA is less affected by the signal noise because CA clusters form as a function of all the observations.

Simulated lidar RCS profiles \( [\text{RCS}(z)] \) are randomly noised \( [\text{RCS}^{\text{noised}}(z)] \) by using the following expression:

\[
\text{RCS}^{\text{noised}}(z) = \text{RCS}(z) + [\alpha \times \chi(z)],
\]

where \( \chi(z) \) is the random noise function taking values between 0 and 1, \( z \) is the height, and \( \alpha \) is a varying parameter as introduced in Eq. (11) to produce different levels of noise. Figure 5 shows the ABL top height as
FIG. 5. Idealized lidar RCS profiles for three noise level cases: (a) $\alpha = 0\%$, (c) $5\%$, and (e) $7\%$, with (b),(d),(f) the results obtained, respectively, by using the CA method, where two clusters are differentiated, as shown by black and white points, respectively, divided by a gray-shaded band.
determined for the same profile $RCS(z)$ affected by three levels of noise, respectively.

Noise level 1 is represented with a value of $\alpha = 0$—that is, $RCS^{\text{noised}}(z) = RCS(z)$—as shown in Fig. 5a. In this case, CA results indicate the ABL top height is obtained at 1097 m (see Fig. 5b). Selected noise level 2 corresponds to an $\alpha$ value of 5%, as shown in Fig. 5c. In this case, the ABL height estimated by CA (see Fig. 5d) is the same as that obtained for the level 1 case ($\alpha = 0$). Changes in clusters appear to be noticeable for $\alpha$ values larger than 6%. For instance, a value of $\alpha = 7\%$ is introduced in Eq. (11) (noise level 3, see Fig. 5e), and CA results are shown in Fig. 5f. Despite that differences in the ABL top height are found in the noise level 3 case, they are lower than 1% in respect to that estimated for the case with $\alpha = 0$ (noise level 1 case).

In summary, these results indicate that the degree of estimation of the ABL top height by applying CA is weakly affected by the signal noise. In fact, while a few methods, such as the derivative methods, estimate the ABL height depending on the value of the RCS gradient in a given point, CA determines the ABL height by taking into account the overall set of observations of a given profile, thus decreasing the dependence of the method on the RCS value in a given point.

c. Effects of cloud contamination

Because of CA dependence on the distribution of the observations, the presence of clouds in the FT near the ABL top must be examined for ABL height estimation. Clouds are characterized by a steep increase of the lidar RCS at the cloud base followed by a strong decrease of the signal. This study is focused on two scenarios, where clouds report high and low values of lidar RCS, respectively.

The presence of clouds with high RCS values is identified by using CA as a cluster change in height at the cloud base, located above that observed for the ABL height estimation, followed by another one at the top. For instance, in the case when two clusters are selected to estimate the ABL height, a first change from cluster 1 to cluster 2 must be shown at the ABL top point. Beyond ABL, a second shift from cluster 2 to cluster 1 at the cloud base is observed, followed by another one from cluster 1 to cluster 2 at the top of the cloud. In this case, similar values corresponding to the cluster observed at ABL heights are found between both these cluster changes, that is, where the cloud is present. For a better understanding of that cloud identification process, an idealized lidar RCS profile is shown in Fig. 6a, where a cloud with high RCS values is localized at around 1460 m. CA results are shown in Fig. 6b, where two clusters are selected and three cluster changes are observed. The first change corresponds to the ABL top height, the second one to the cloud base, and the last one to the cloud top. In these cases, clouds are detected by using CA when two consecutive cluster changes are observed (above the cluster change corresponding to the ABL top). Differences in ABL height are calculated for those cases without and with cloud presence (high RCS values) near the ABL top, obtaining variations in ABL height estimation lower than 1.8%. Therefore, in the case when clouds are detected in the lidar RCS profile, height points where a cluster change is observed below the cloud base are considered to be the ABL top heights.

In the case of clouds present with low RCS values, a cluster change is unobserved along the profile above the ABL top height. An idealized lidar RCS profile is shown in Fig. 6c, where a cloud is localized at around 1400–1500 m with low RCS. CA results are unaltered by the low RCS cloud presence, as shown in Fig. 6d. In this case, cloud contamination presents an insignificant effect in ABL top height estimation.

d. Real case studies

A first application of the cluster analysis to real lidar measurements in order to compare the ABL top height as estimated by different methods was recently performed and preliminary results were presented in Toledo et al. (2012). A more detailed description of the CA method performance as applied to lidar measurements has been presented in this work (see previous sections). Next, real case studies are examined in order to evaluate the reliability of the CA method to retrieve the ABL top height from real lidar measurements under different atmospheric conditions.

Lidar RCS profiles are averaged and with its variance are introduced in the CA algorithm. By using the K-means method, the process starts only with two clusters, and both the $D$ and DB indexes are calculated. Then, the process is repeated with one more cluster included, and the new values of the $D$ and DB are calculated again. This process continues to include more clusters until both the $D$ index decreases and the DB index increases in comparison with the previous values, indicating in that case that the optimal number of clusters is obtained. In the case of when the hierarchical method is applied to lidar measurements, the number of cluster is decided on the basis of the dendogram explained in section 2 (see Fig. 2). In both those cases, the ABL top is identified at the height point where the first change from a cluster to another one is observed.

Lidar observations carried out at the atmospheric sounding station El Arenosillo (ARN) of the Instituto Nacional de Técnica Aeroespacial (INTA; 37.1°N, 6.7°W; 40 m msl) in the framework of the Atmospheric
Minor Species Relevant to Ozone Chemistry at Both Sides of the Subtropical Jet (AMISOC) project have been used for these case studies. The INTA-ARN station is a rural background environment located at the southwestern Iberian Peninsula and around 1 km from the coastline of the Atlantic Ocean. It is generally characterized by stable atmospheric conditions, only disturbed when dust intrusions from the Sahara Desert and other occasional aerosol events occur. The lidar system used is a MicroPulse Lidar version 3 (MPL-3), an eye-safe commercially available backscatter lidar, consisting mainly of a single-wavelength (523 nm), high-repetition frequency (2500 Hz), and low-power (maximum 7 μJ) laser. More instrumental details are found in Campbell et al. (2002). MPL-3 was operating in full-time unattended mode with a vertical resolution of 15 m and a temporal integration time of 1 min. RCS profiles were averaged over 10 min for both variance calculation and signal-to-noise ratio (SNR) enhancement, being the computing time required to process a full day (144 profiles of 10-min average) lower than 5 min.

To study the reliability of the CA method to retrieve the ABL height under different conditions, results are shown for three specific cases: 1) the ABL evolution

---

**FIG. 6.** Idealized lidar RCS profiles for (a) high and (c) low RCS values at altitudes 1400–1500 m corresponding to cloud occurrence, with (b),(d) results obtained, respectively, by using the CA method, where two clusters are differentiated, as shown by solid and dashed lines, respectively, divided by a gray-shaded band.
along the convective period; 2) the ABL variation during an afternoon transition, where a residual layer is observed above; and 3) the morning ABL transition with a multilayered aerosol scenario. These three cases were observed during the AMISOC-ARN campaign, and they are representative of the usual atmospheric conditions present during the day in the study area and not related to specific dusty scenarios or another kind of particular aerosol events occurring during that measurement period. Otherwise, the complexity of the ABL dynamics under particular aerosol conditions must be studied in much more detail, which is out of the scope of this work.

Temporal evolution of the lidar RCS registered from 1030 to 1700 UTC 26 May 2012 is shown in Fig. 7a, regarded under convective conditions. It must be mentioned, since time periods are referred to UTC instead of local time in this work, that sunlight ranges from the sunrise observed at around 0530 UTC to the sunset at around 1915 UTC. This time interval would correspond to the convective period, more affected by the solar activity. As stated before, two clusters were selected on the basis of the $D$ and DB validity indexes for ABL height estimation, identifying the ABL top as the height point where a cluster change is observed. In this period, an increase of the ABL top height is observed from 1030 to 1230 UTC 26 May, reaching a maximal value of 1634-m height at 1530 UTC (see Fig. 7a). Individual 10-min averaged RCS profiles for three selected times—1150, 1350, and 1640 UTC—are shown in Figs. 7b–d, respectively, with their corresponding variance profiles as shown in Figs. 7e–g, respectively. The ABL top height as

![Graph showing temporal evolution of lidar RCS and ABL top height](image)
calculated by using CA is also indicated in each case by a horizontal solid line, representing the height point where two clusters are differentiated (black and white circles, respectively). ABL top values of 1176-, 1416-, and 1356-m height are found for each case. As expected, the highest ABL top height among these selected profiles was reached at around 1400 UTC (see Fig. 7c) under more convective atmospheric conditions. In addition, those ABL top height values estimated are also confirmed by the maximum observed in the corresponding variance profiles (see Figs. 7e–g) at the height ranges where the cluster change can be also found.

A second case, representing the ABL afternoon transition, is shown in Fig. 8a, corresponding to the lidar RCS registered from 1700 to 2130 UTC 3 June 2012. A residual layer can be observed above the ABL top height gradually spreading along this time period (as indicated by white arrows in Fig. 8a). Moreover, the nocturnal layer develops into a stable boundary layer, whose top height is progressively decreasing as atmospheric convective conditions are vanishing. As expected, the maximal and minimum ABL height points are reached at 1700 and 2130 UTC during this period, with ABL height values of about 500 and 200 m, respectively. Individual 10-min averaged RCS profiles at 1710, 1810, and 1910 UTC are shown in Figs. 8b–d, respectively, where each symbol represents the altitude range associated with different clusters. The corresponding profiles of the variance are shown in Figs. 8e–g, respectively. The ABL top height, as determined by the
CA method, is hence identified at the height point where the first two clusters are differentiated (as represented by stars and circles, respectively, in Figs. 8–d). Particular ABL top values of 465-, 450-, and 435-m height are found for each case, as indicated by a horizontal solid line (see Figs. 8b–d). In this “afternoon transition” case, unlike the previous convective period (see Fig. 7), four to five clusters were required to be introduced in the CA algorithm according to the cluster validation criteria. However, despite a larger number of clusters being needed due to the presence of that elevated residual layer presenting a multilayered structure, an accurate ABL top height estimation was indeed obtained. This fact can be also observed in the variance profiles, where several maxima are found. In particular, the variance profile at 1810 UTC (see Fig. 8f) shows a clear maximum between 400- and 500-m height, unlike the other two plots (Figs. 8e and 8g) where several maxima are found, and the ABL top height can be ambiguously identified. In these situations, the vertical distribution of aerosols is used by CA to properly determine the ABL height. Therefore, this case illustrates the importance of cluster validation in order to estimate the number of clusters required for discrimination between the ABL and other residual layers observed above.

The third case corresponding to the early morning evolution of the lidar RCS registered from 0420 to 0650 UTC 6 June 2012 is shown in Fig. 9a, where a multilayer aerosol scenario is observed. Regarding lidar RCS values, a clear three-layered structure is identified.
during the first 2 h of this period (see Fig. 9a), with the sublayers ranging from the ground to 150-m height, from 270- to 360-m height, and from 400- to 1200-m height, respectively. The lower layer corresponds to the ABL, while the higher ones are aerosol layers (as indicated by white arrows in Fig. 9a). The middle layer completely vanished after around 0500 UTC, whereas the upper layer was present until the end of this period, but its thickness is decreasing as the solar activity is increasing in the morning. As in the afternoon transition case, four to five clusters were needed to retrieve the ABL top height. To illustrate the clustering structure, individual 10-min averaged RCS profiles at 0510, 0600, and 0620 UTC are shown in Figs. 9b–d, respectively, with their corresponding variance profiles, also shown in Figs. 9e–g. As before, the ABL top height obtained by using CA is indicated by a solid line, and the different symbols represent each cluster considered (Figs. 9b–d). In particular, the first two variance plots (see Figs. 9e and 9f) show maximums well defined, corresponding to those height ranges where the ABL top height was estimated (see the horizontal line in Figs. 9b and 9c). In addition, the importance of the variance profiles is highlighted for the first case (see Figs. 9a and 9e), since negative gradients are found in the lidar RCS profiles below the ABL top height. Hence, despite a clear gradient in lidar signal is weakly found at ABL/FT transition altitudes, CA can estimate the ABL top height by discriminating clearly different clusters, which would correspond to particular layers present in the atmosphere.

In summary, all these cases show the good performance of the CA method in retrieving the ABL top height under both stable conditions and the presence of residual aerosol layers above being dependent on the number of clusters required, according to the optimal validity indexes considered. Hence, the complexity of the atmosphere evaluated in terms of discriminated aerosol layers can be examined by the CA method as applied to lidar measurements.

The evaluation of the CA approach in comparison with other usually used methods for ABL top height determination is ongoing. These all methods will be also applied to those recently performed lidar measurements at the INTA-ARN station in the frame of the AMISOC project. In addition, the validation of all these methods against standard radiosoundings as a reference will complete that study.

4. Conclusions

The cluster analysis (CA) approach has been applied for the first time, to our knowledge, to lidar measurements in order to calculate the ABL top height. The method is based on the information provided from the vertical aerosol distribution and variation of their concentration (variance). Results obtained show that the ABL height can be estimated by using the CA approach under a wide variety of situations, assuming an aerosol exchange between the surface and FT. In addition, by comparison with other techniques, it has been found that CA is able to retrieve a unique ABL height in situations when the methods based on both the vertical distribution of the aerosol concentration and the statistical analysis present an uncertainty in their ABL top height estimation.

In relation to the robustness of the CA approach, the effect of the lidar RCS noise in determining the ABL top height has been analyzed. Unlike derivative methods, CA results are slightly affected by lidar signal noise. ABL top heights as obtained for “noised” lidar RCS with values of $\alpha < 6\%$ is the same as those found for the noise level 1 ($\alpha = 0$), and only small differences ($< 1\%$) in the ABL top height estimation are observed for “noisier” RCS profiles ($\alpha = 7\%$).

The efficiency of the CA approach under cloudy conditions near the ABL top has been also tested. Differences in ABL top height as estimated between cloudy and clear skies were found to be lower than 1.8% for a high RCS signal, while no effect was observed for low RCS cloud conditions. Therefore, the CA approach is able to accurately obtain the ABL top height when lidar signals are affected by cloud presence.

Finally, CA performance on ABL top height determination for three different case studies has been also presented to illustrate the application of this new method to real lidar measurements as well as its reliability to retrieve the ABL height under different atmospheric conditions. Nonhierarchical CA results indicate that the evolution of the ABL height is well reproduced by applying this method to lidar measurements performed during the day (the same results are obtained by applying hierarchical CA; data not shown).

This work mainly presents the cluster analysis as a reliable method to be applied to lidar measurements for ABL studies. The evaluation of the CA approach in comparison with other usually used methods for ABL top height determination is ongoing.

Acknowledgments. This work was supported by the Spanish Ministerio de Economía y Competitividad (MINECO) under Grant CGL2011-24891 (AMISOC project). The authors specially thank the INTA-Atmospheric Sounding Station “El Arenosillo” (ARN) staff for their valuable assistance and support in the lidar system maintenance.
REFERENCES


