To cite this version : Rieutord, Thomas. *Automatic detection of the boundary layer height using Doppler lidar measurments.* (2014) [Mémoire]

Automatic detection of the boundary layer height using Doppler lidar measurements

Thomas Rieutord

École Nationale de la Météorologie

Master degree thesis

Acknowledgments

Ces remerciements ne seraient pas complets s'ils ne traversaient pas l'Altlantique. Merci à tous ceux qui, de près ou de loin, ont participé à faire de stage ce qu'il est. Je commencerais par remercier Alain Dabas, sans qui je n'aurais pas pu obtenir ce stage. De l'autre coté de l'océan, mes pensées vont d'abord à mes hôtes Alan Brewer et Mike Hardesty, ainsi que le reste de l'équipe, qui m'ont fait découvrir les lidars et les "jours typiques". Merci en particulier Aditya Choukulkar pour son aide précieuse au travail et sa gentillesse sans bornes. Je remercie aussi Christoph Senff pour ses conseils avisés et avenants. Enfin, je remercie chaleureusement mes compatriotes et amis Antoine, Aurélien et Guillaume, ainsi que Cyrus et Thomas, pour leur savoir-vivre et leurs cours d'anglais.

These acknowledgments will not be complete if they don't cross the Atlantic. Thanks to everyone who have any implication to make this internship what it is. I will start with Alain Dabas, who made this internship possible. On the other side of the ocean, my thoughts go first to my hosts Alan Brewer and Mike Hardesty, as well as the entire team, for let me discover lidars and "typical days". Special thanks for Aditya Choukulkar, whose help was precious at work, and for his boundless kindness. I would like to thank Christoph Senff for his wise and welcome advice. Finally, I warmly thank my fellow countrymen and friends Antoine, Aurélien and Guillaume, as well as Cyrus and Thomas, for their friendliness and their English lessons.

Abstract

L'un des premiers paramètres qui vient à l'esprit lorsque l'on veut décrire la couche limite atmosphérique est sa hauteur. En effet, lorsqu'on explique par exemple le cycle diurne de la couche limite, on le met généralement en image avec la variation de la hauteur de couche limite. Cela en fait un paramètre très important dans tous les domaines de recherche sur la couche limite. Il est notamment très utile pour étudier la qualité de l'air et la paramétrisation de la turbulence, entre autres.

Malheureusement, l'accès à la hauteur de couche limite par la mesure n'est pas chose aisée. En effet, la contrepartie d'une utilité dans des domaines variés est une variabilité complexe, mêlant tous ces domaines. L'œil humain peut aisément distinguer la transition entre couche limite et atmosphère libre sur des profils de turbulence et d'intensité rétrodiffusée, mais c'est beaucoup plus difficile pour un ordinateur, car ces profils présentent des caractéristiques très différentes en fonction de la situation. L'objectif de ce stage est de construire un algorithme capable de tirer profit de toute l'information fournie par les lidars Doppler pour fonctionner en toute situation. Deux types de méthodes ont été testés ici : une première basée sur la détection de pics dans un profil, une deuxième basée sur de la classification de données.

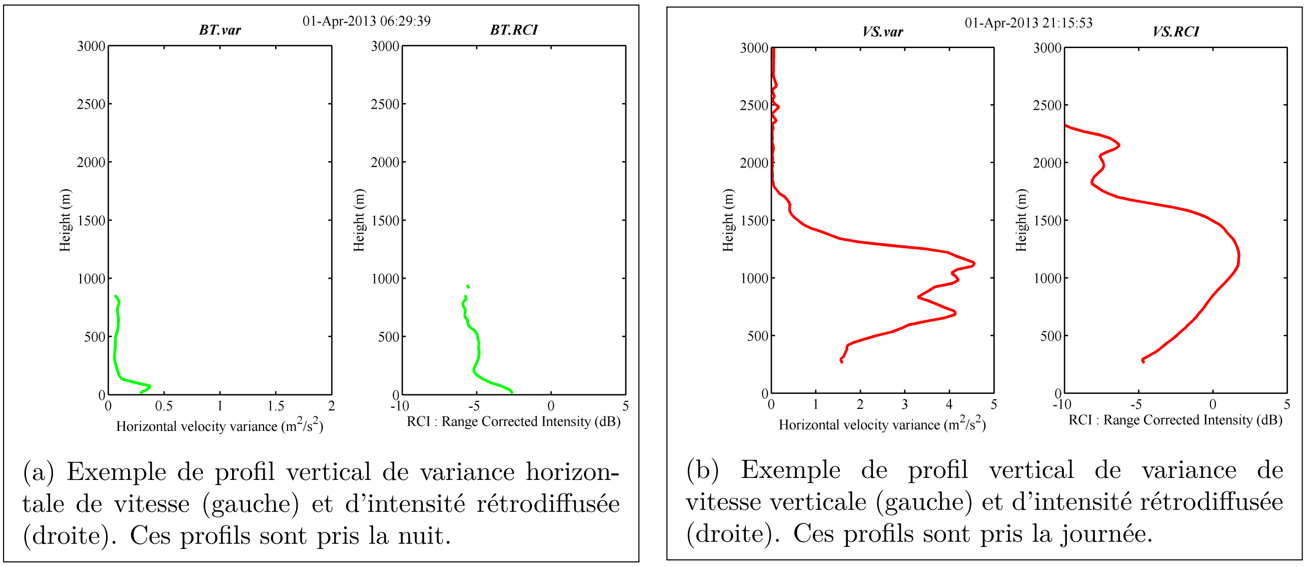
Dans la première méthode, l'idée est de transformer le profil mesuré (profil de turbulence, d'intensité rétrodiffusée) de façon à faire apparaitre la transition entre couche limite et atmosphère libre comme un pic (un exemple de transformation est le gradient vertical). On choisit ensuite les pics en cherchant une continuité avec les profils voisins. Sur certains types de profils, comme ceux de turbulence, on applique une méthode de seuillage pour lequel le seuil est déterminé automatiquement en fonction du profil. Dans la deuxième méthode, l'idée est d'agglomérer les points du profil en groupes homogènes : un groupe pour la couche limite, un autre pour l'atmosphère libre. La distance utilisée pour quantifier la dissimilarité entre les points du profil se base sur les données mesurées par le lidar (turbulence, intensité rétrodiffusée). Les groupes sont créés par l'algorithme des "K-means". Ensuite on définit la hauteur de couche limite par la hauteur on l'on change de groupe.

Résumé long

L'élément descriptif essentiel de la couche limite atmosphérique est sa hauteur. En effet, lorsqu'on s'intéresse par exemple au cycle diurne de la couche limite, on le met généralement en image avec la variation de la hauteur de couche limite. Cela en fait un paramètre très important dans tous les domaines de recherche sur la couche limite. Il est notamment très utile pour étudier la qualité de l'air (puisque les aérosols émis au sol vont rester concentrés dans cette couche-là), la paramétrisation de la turbulence (puisqu'il délimite le domaine où la turbulence est importante), entre-autres.

Bien que l'accès à la hauteur de couche limite par la mesure ne soit pas immédiat, il est possible de détecter la hauteur de couche limite à partir d'observations. Les profils verticaux de température potentielle sont marqués d'une inversion au sommet de la couche limite ; les profils de concentration en aérosols présentent un fort gradient, de même que les profils de turbulence (exemple sur les figures 1a et 1b). Les lidars Doppler mesurent en continu des profils de vent (d'où l'on tire de l'information sur la turbulence) et d'intensité rétrodiffusée (proportionnelle au contenu en aérosol) avec une haute résolution verticale et temporelle. Cela en fait un instrument très approprié pour la détection en temps réel de la hauteur de couche limite.

Le lidar qui a fourni les données pour ce travail effectue 3 scans, avec une résolution temporelle finale de 20 minutes. Le premier scan est conique, répété pour 3 angles d'élévations. Il fournit des mesures de vent horizontal (force et direction). Le deuxième parcourt les angles d'élévations de 0 à 24° pour 2 azimuts orthogonaux. Il est appelé « nœud papillon » et convient particulièrement à la mesure des basses couches. Le cycle de 20 minutes est complété en pointant vers le zénith.



(a) Exemple de profil vertical de variance horizontale de vitesse (gauche) et d'intensité rétrodiffusée (droite). Ces profils sont pris la nuit.

(b) Exemple de profil vertical de variance de vitesse verticale (gauche) et d'intensité rétrodiffusée (droite). Ces profils sont pris la journée.

Figure 1: Example de profils mesurés par le lidar Doppler.

Cependant, l'automatisation du calcul de la hauteur de couche limite à partir d'observations, si pertinentes soient-elles, est toujours un défi d'actualité. L'œil humain peut aisément distinguer la transition entre couche limite et atmosphère libre sur des profils de turbulence et d'intensité rétrodiffusée, mais c'est beaucoup plus difficile pour un ordinateur car ces profils présentent des caractéristiques très différentes en fonction de la situation. Beaucoup d'algorithmes ont prouvé leur efficacité sur une quantité plus ou moins restreinte de conditions atmosphériques. L'objectif de ce stage est de construire un algorithme capable de tirer profit de toute l'information fournie par les lidars Doppler pour fonctionner en toute situation. Deux types de méthodes ont été développées : une première basée sur la détection de pics dans un profil vertical, une deuxième basée sur de la classification de données.

La première méthode utilise la détection de « pics » (maxima locaux). Son avantage est la variété de phénomènes pouvant être détectés par des maxima locaux. Pour les profils de types « escalier » comme sur les figures 1a et 1b, le maximum intéressant va être celui du gradient. Mais on peut aussi penser au jet de basse couche dans un profil de vent. Deux techniques de détection de la hauteur de couche limite à partir de pics vont être détaillées, car ce sont les plus utilisées.

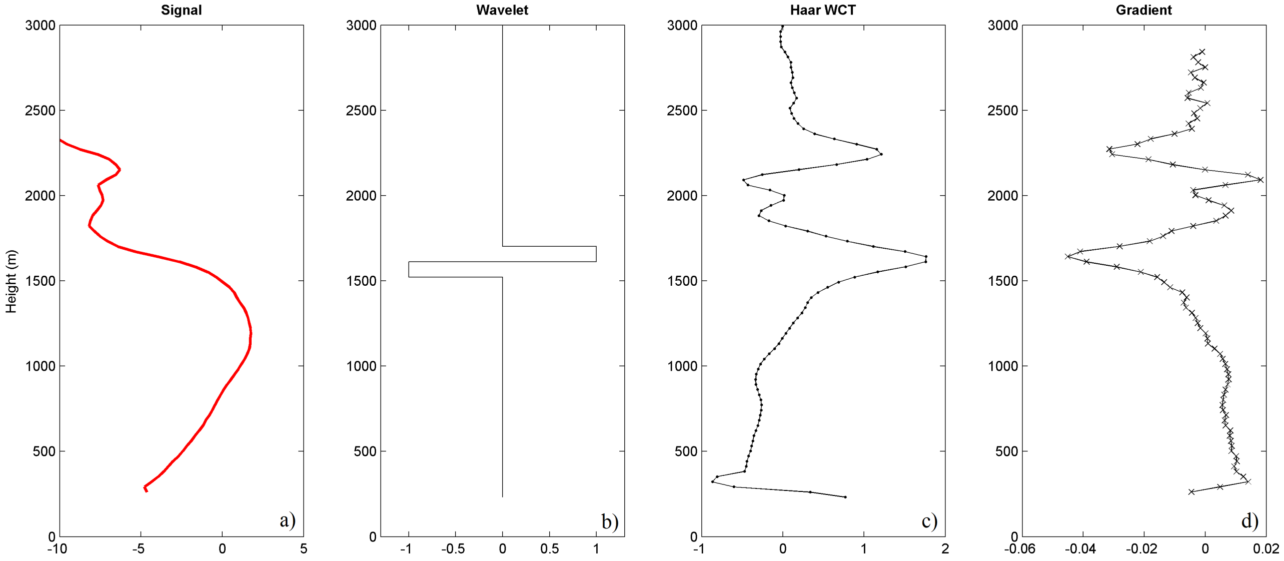
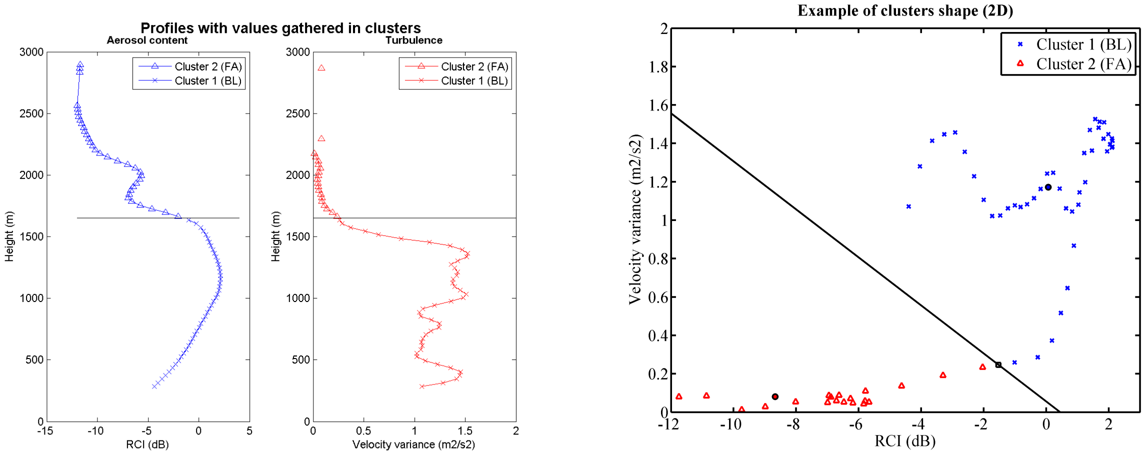


Figure 2: Exemple de transformée en ondelette de Haar d'un profil d'intensité rétrodiffusée et comparaison avec le gradient. (a) Profil original. (b) Ondelette de Haar. (c) Transformée en ondelette de Haar du profil original. (d) Gradient du profil original.

Pour la première technique, prenons par exemple un profil d'intensité rétrodiffusée (figure 1b, droite) : la transition entre couche limite et atmosphère libre est marquée par un palier dans le profil. Les aérosols sont plus concentrés dans la couche limite que dans l'atmosphère libre. Si l'on prend le gradient vertical de l'intensité rétrodiffusée, le profil présentera un pic au niveau de la transition, comme on peut le voir sur la figure 2d. La figure 2 montre un profil vertical d'intensité rétrodiffusée (a), une ondelette de Haar (b), la transformée en ondelettes du profil (c), et le gradient classique de de profil. Pour être moins sensible au bruit, nous avons préféré au gradient classique une transformation en ondelette de Haar. La transformée en ondelette de Haar consiste à calculer la convolution de notre profil avec une ondelette de Haar (fonction escalier, fig. 2b). Le profil ainsi obtenu présente des hautes valeurs lorsque le profil original (d'intensité rétrodiffusée) "ressemble" à une fonction en escalier. Le pic qui correspond bien à la transition entre la couche limite et l'atmosphère libre est généralement le plus intense. De plus, les pics dans les profils voisins en temps peuvent donner une information sur la continuité qui peut être exploitée.

Pour illustrer la seconde technique, prenons par exemple un profil de variance de vitesse verticale (figure 1b, gauche). Cette quantité est un traceur de la turbulence, caractéristique de la couche limite. On peut remarquer que le profil lui-même présente un plus grand nombre de maxima locaux que le profil d'intensité rétrodiffusée. La recherche du maximum de gradient peut être perturbée par ces variations. Il est donc plus judicieux de caractériser le sommet de la couche limite par l'altitude à laquelle la turbulence disparait, c'est-à-dire l'altitude la plus élevée en dessous de laquelle la variance garde des valeurs suffisamment hautes. La valeur minimum de variance de vitesse pour que le profil soit considéré comme non-turbulent est déterminée automatiquement en fonction du profil.

La seconde méthode développée utilise un algorithme de classification de données. Elle se base sur l'hypothèse que les valeurs de turbulence et de concentration en aérosol sont significativement plus élevées dans la couche limite que dans l'atmosphère libre. En considérant un profil vertical où chaque point de mesure contient plusieurs informations (variance de la vitesse, intensité rétrodiffusée), on regroupe les points de mesures qui ont des informations similaires avec l'algorithme des "K-means". La figure 3a montre deux profils pour lesquels on va effectuer cette opération (profil d'intensité rétrodiffusée à gauche, profil de variance de vitesse verticale à droite, pris de jour). Les points contenant les deux informations sont représentés sur la figure 3b. Chaque point est caractérisé par une valeur d'intensité rétro-diffusée (abscisse) et une valeur de variance de vitesse (ordonée). On rassemble ces points de façon à identifier les groupes « couche limite » (croix bleues) et « atmosphère libre » (triangles rouges). La hauteur de couche limite est alors définie par la hauteur où l'on change de groupe. Le nombre de groupe est donné a priori. Pour chaque groupe, on définit un centroïde : le point moyen du groupe, sans réalité physique. Un point appartient à un groupe s'il est plus près du centroïde de ce groupe là que des autres centroïdes. L'algorithme construit les groupes de façon itérative, en faisant bouger les centroïdes.



(a) Ces deux profils sont rassemblés en un seul profil dont les points de mesures contiennent les deux informations.

(b) Les points du profil forment des groupes que l'on identifie comme la couche limite (triangles rouges), et l'atmosphère libre (croix bleues).

Figure 3: Méthode de classification appliquée à un profil réel.

Cette méthode a été enrichie tardivement d'une variante permettant d'évaluer l'incertitude de l'estimation. Il s'agit de « classification floue » (fuzzy clustering). L'algorithme est inchangé, l'innovation vient du fait que les points appartiennent partiellement à tous les groupes au lieu d'un seul. Cela permet d'avoir une représentation continue de nos groupes, d'où l'on peut potentiellement tirer une information sur l'incertitude de l'estimation. Le potentiel de cette méthode n'a pas pu être exploré pendant le temps imparti.

Ces méthodes ont été testées sur une base données de 21 jours, avec un profil de chaque donnée toutes les 20 minutes. Les résultats détaillés sont présentés sous la forme d'une étude de cas. Les points forts et les limites de chaque méthode sont expliqués sur un jour particulier. Il en ressort que les estimateurs utilisant des données du scan "nœud papillon" ne sont pas en mesure de détecter le sommet de la couche limite convective à cause de sa portée limitée. De façon symétrique, les estimateurs utilisant les données venant du pointage vers le zénith ne sont pas en mesure de détecter la hauteur de couche limite la nuit à cause de la portée minimum du lidar. Les mesures de vent horizontal apparaissent n'être pertinentes que lors de conditions météorologiques spécifiques (jet de basses couches, cisaillement au niveau de l'inversion).

Les méthodes sont ensuite évaluées sur l'ensemble de la base de données par deux examinations visuelles indépendantes. La figure 4 montre les résultats d'une de ces deux examinations. Chaque barre verticale correspond à un estimateur, la partie bleue représente la proportion de profils pour lesquels l'estimateur donne la hauteur identifiée sur les données comme étant la hauteur de couche limite. La partie jaune correspond à la proportion de profil où l'estimateur n'a pas suffisamment de données pour

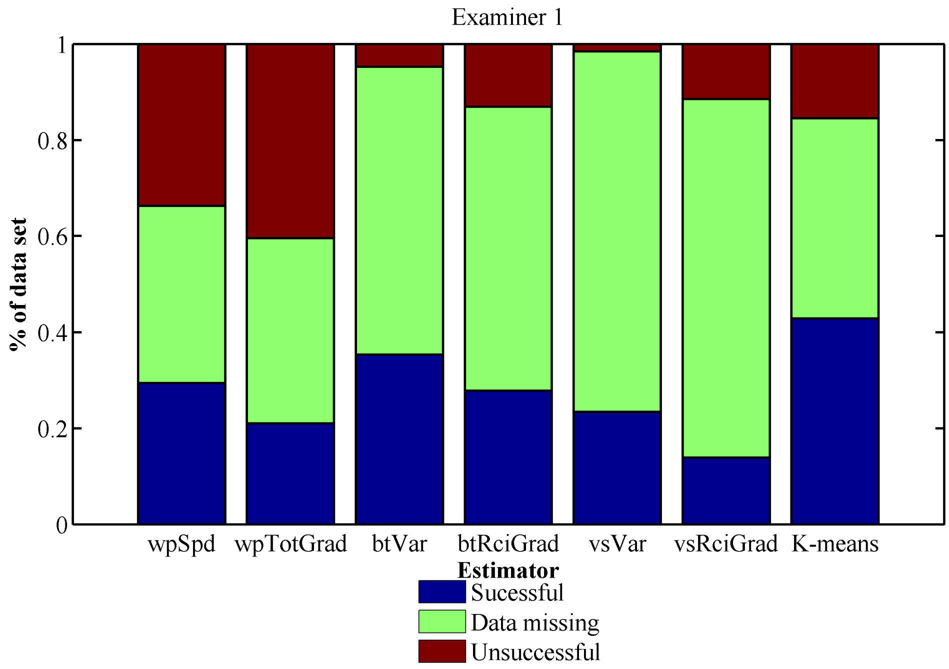


Figure 4: Résultats de l'examen visuel des estimateurs. En abscisse sont les estimateurs :wpSpd et wpTotGrad utilisent le vent horizontal, btVar et vsVar utilisent le seuillage, btRciGrad et vsRciGrad utilisent la transformée en ondelettes. Pour chaque estimateur, la base de donnée est découpée en trois catégories : Estimation réussie (bleu), Donnée insuffisante (jaune), Estimation ratée (rouge).

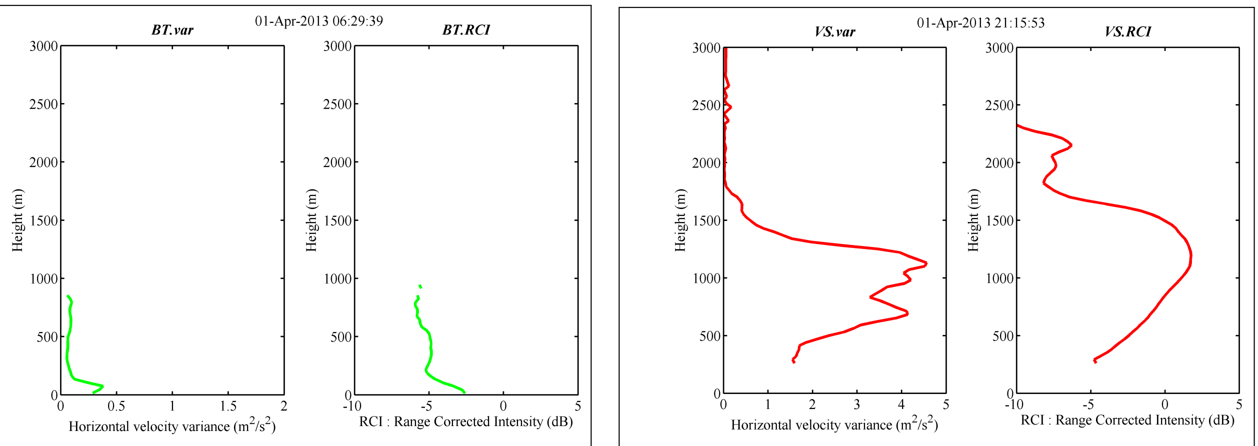
fonctionner. La partie rouge représente la proportion de profil où l'estimateur devrait fonctionner mais ne fonctionne pas. Les résultats montrent que la méthode de seuillage est la plus fiable, avec un taux de réussite proche de 80%, mais c'est aussi la méthode la moins disponible avec respectivement 25% et 42% de disponibilité pour les scans zénith et « nœud papillon ». La méthode la plus disponible est celle utilisant le vent horizontal avec 68% de disponibilité, mais c'est aussi la moins efficace, avec un taux de réussite de l'ordre de 30%. La méthode de transformée en ondelettes de Haar se situe entre les deux précédentes, avec autant de disponibilité que la méthode de seuillage, mais un taux de réussite d'environ 50%. Enfin, la méthode de classification de donnée fait figure de bon compris avec un taux de réussite de 60% (2e plus élevé), et une disponibilité de 62% (2e plus élevée).

Extended abstract

An essential description of the atmospheric boundary layer element is its height. Indeed, for example when considering the diurnal cycle of the boundary layer, one generally uses the image with the variation of the boundary layer height. This makes it a very important parameter in all areas of research on the boundary layer. It is particularly useful for studying the quality of the air (since the emitted aerosol ground will remain concentrated in the layer one), the parametrization of the turbulence (since it defines the area where turbulence is important).

Although access to the boundary layer height from measurements is not immediate, it is possible to detect the boundary layer height from observations. Vertical profiles of potential temperature are marked with an inversion at the top of the boundary layer; the aerosol concentration profiles show a strong gradient, as well as the turbulence profiles (example in figures 5a and 5b). Doppler lidar continuously measure wind profiles (which is similar to turbulence) and back-scattered intensity (proportional to the aerosol) with high vertical and temporal resolution. This makes it very suitable for real-time detection of the boundary layer height.

The lidar which provided the data for this work makes three scans, with a final time resolution of 20 minutes. The first scan is conical, repeated for three angles of elevations. It provides measurements of horizontal wind (speed and direction). The second elevation angles travels 0 to 24° for two orthogonal azimuths. It is called "bowtie" and is particularly suitable for measuring low levels. The 20-minute cycle is completed by staring vertically.



(a) Example of vertical profile of horizontal velocity variance (left) and back-scattered intensity (right). These profiles are taken at night.

(b) Example of vertical profile of vertical velocity variance (left) and back-scattered intensity (right). These profiles are taken during the day.

Figure 5: Example of profiles measured by Doppler lidar.

However, automated calculation of the boundary layer height from observations, if relevant they are, is always a current challenge. The human eye can easily distinguish the transition between limit and free atmosphere turbulence profiles and back-scattered intensity layer, but it is much more difficult for a computer because these profiles have very different characteristics depending on the situation. Many algorithms have proven their effectiveness regarding more or less restricted atmospheric conditions. The goal of this internship is to build an algorithm able to take advantage of any information provided by the

Doppler lidar to work in any situation. Two types of methods have been developed: one based on the detection of peaks in a vertical profile, the second is based on data clustering.

The first method uses the detection of peaks (local maxima). Its advantage is the variety of phenomena that can be detected. For profiles of type "staircase" as in figures 5a and 5b, the maximum searched for will be the one calculated from the gradient. But we can also think of the low-level jet in a wind profile. Two of the most used techniques for detecting the boundary layer height based on peaks detection will be detailed.

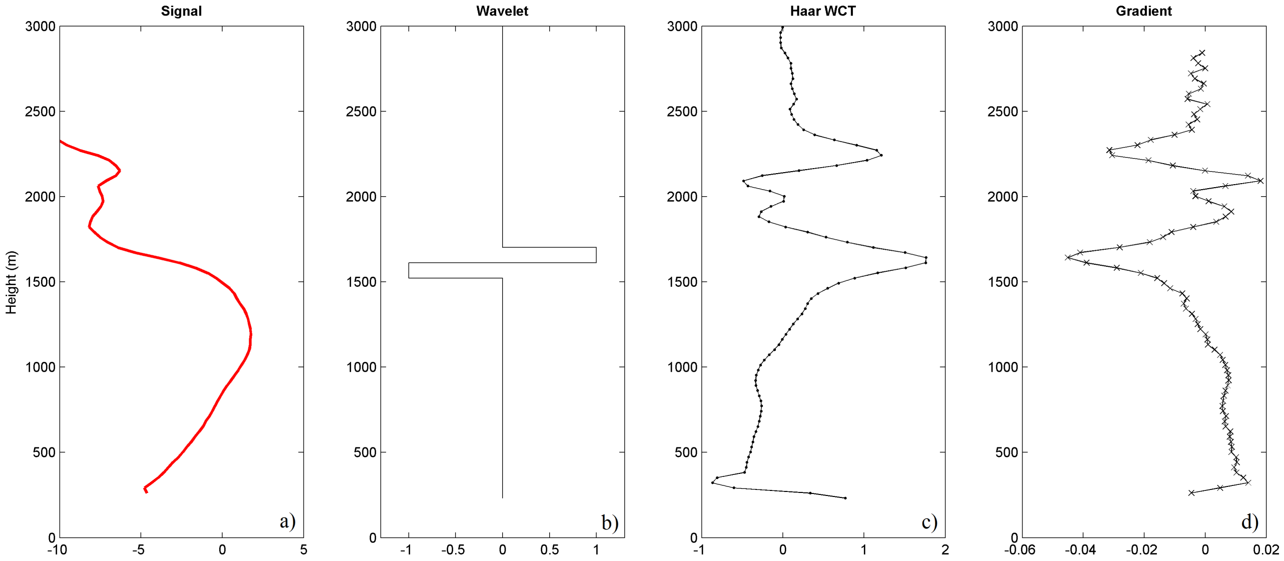
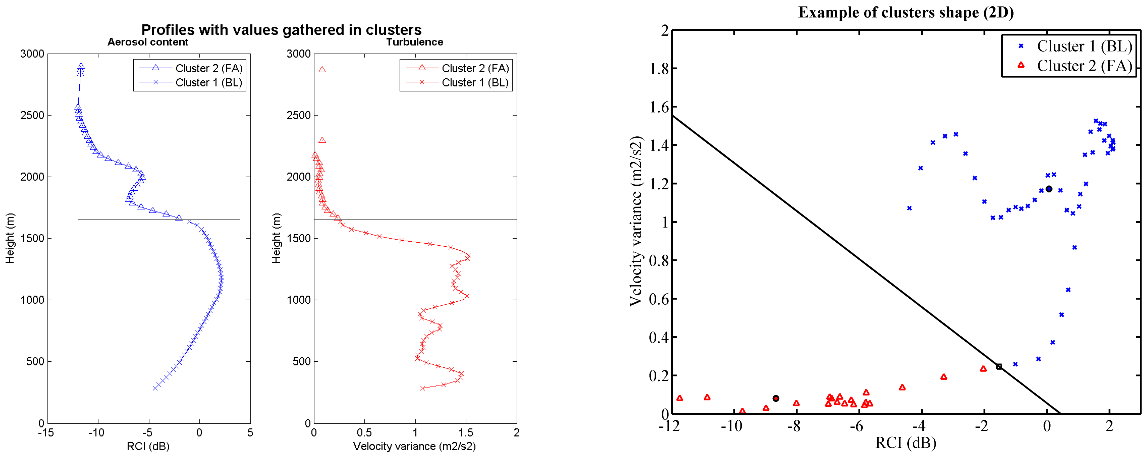


Figure 6: Example of a Haar wavelet transform of a back-scattered intensity and comparison with the gradient profile. (a) The original profile. (b) Haar Wavelet. (c) Haar wavelet transform in the original profile. (d) Gradient of the original profile.

For the first technique, take for example a back-scattered intensity profile (figure 5b, right): the transition between boundary layer and free atmosphere is characterized by a plateau in the profile. Aerosols are more concentrated in the boundary layer than in the free atmosphere. Taking the vertical gradient of the back-scattered intensity, the profile will have a peak at the transition, as shown in figure 6d. Figure 6 shows a vertical profile of back-scattered intensity (a), a Haar wavelet (b), the wavelet transform of the profile (c), the conventional gradient and profile. To be less sensitive to noise, the Haar wavelet transform has been preferred to the classical gradient. The Haar wavelet transform is to calculate the convolution of this profile with a Haar wavelet (staircase function, fig. 6b). The profile thus obtained has high values when the original profile (back-scattered intensity) "looks like" a step function. The peak that corresponds to the transition between the boundary layer and the free atmosphere is usually the most intense. Moreover, the peaks in the neighboring time profiles may give information about the continuity that can be exploited.

To illustrate the second technique, we take for example a profile of vertical velocity variance (figure 5b, left). This is a turbulence tracer, a characteristic of the boundary layer. It may be noted that the profile itself has a larger number of local maxima than the back-scattered intensity profile. The search for the maximum gradient can be affected by these changes. It is therefore more appropriate to characterize the top of the boundary layer by the altitude at which the turbulence disappears: the highest altitude below which the variance values is sufficiently high. The minimum variance value for the velocity profile to be considered non-turbulent is determined automatically, depending on the profile.

The second method uses a clustering algorithm. It is based on the assumption that the values of turbulence and aerosol concentration are significantly higher in the boundary layer than in the free atmosphere. Considering a vertical profile where each measurement point contains several information (velocity variance, intensity backscattered), the measurement points that have similar information are grouped by the K-means algorithm. Figure 7a shows two profiles for which we performed this operation (backscattered intensity profile on the left, variance profile on the right vertical speed, profiles were taken during day). The items containing both information are presented in figure 7b. Each point is characterized by a back-scattered intensity value (x-axis) and a velocity variance value (y-axis). These points are combined to identify "boundary layer" (blue crosses) and "free atmosphere" (red triangles) clusters. The boundary layer height is then defined by the height of the clusters border. The cluster number is given *a priori*. For each cluster, we define a centroid: the midpoint of the group, without physical reality. A point belongs to a cluster if it is closer to its centroid than the others. The algorithm build clusters iteratively by moving centroids.



(a) These two profiles are combined into a single profile with the measurement points contain both information.

(b) The profile points form clusters that are identified as the boundary layer (red triangles), and the free atmosphere (blue crosses).

Figure 7: Cluster analysis method applied on a real profile.

This method was later expanded to assess the uncertainty of the estimate. It is called "fuzzy clustering". The algorithm is unchanged, the innovation is that points now belong partially to all clusters instead of one. This allows for a continuous representation of our clusters, which can potentially give information on the uncertainty of the estimate. The potential of this method has not been explored during the allotted time.

These methods were tested on a 21 days data base, with a profile taken every 20 minutes. Detailed results are presented as a case study. The strengths and limitations of each method are explained on this particular day. It shows that the estimators using data from the "bowtie" scan are not able to detect the top of the convective boundary layer due to its limited range. Estimators using data from vertical staring are not able to detect the height of the boundary layer at night because of the minimum range of the lidar. Horizontal wind measurements appear only to be relevant under specific weather conditions (low-level jet, shear at the inversion).

The methods are then assessed on the whole database by two independent visual examinations. Figure 8 shows the results of one of these examinations. Each vertical bar represents an estimator, the blue area represents the proportion of profiles for which the estimator gives the boundary layer height as identified on the data. The yellow area is the proportion of the profile where estimator has not enough data to proceed. The red area represents the proportion of profile where the estimator should work but fails. The results show that the thresholding method is more reliable, with a success rate close to 80%, but it is also the method the least available with 25% and 42% of availability respectively for vertical staring and "bowtie ". The most available method is using the horizontal wind information with 68% availability, but it is also the least effective, with a success rate of around 30%. The Haar wavelet transform method is between the two previous ones, with the same availability as the thresholding method but a success rate

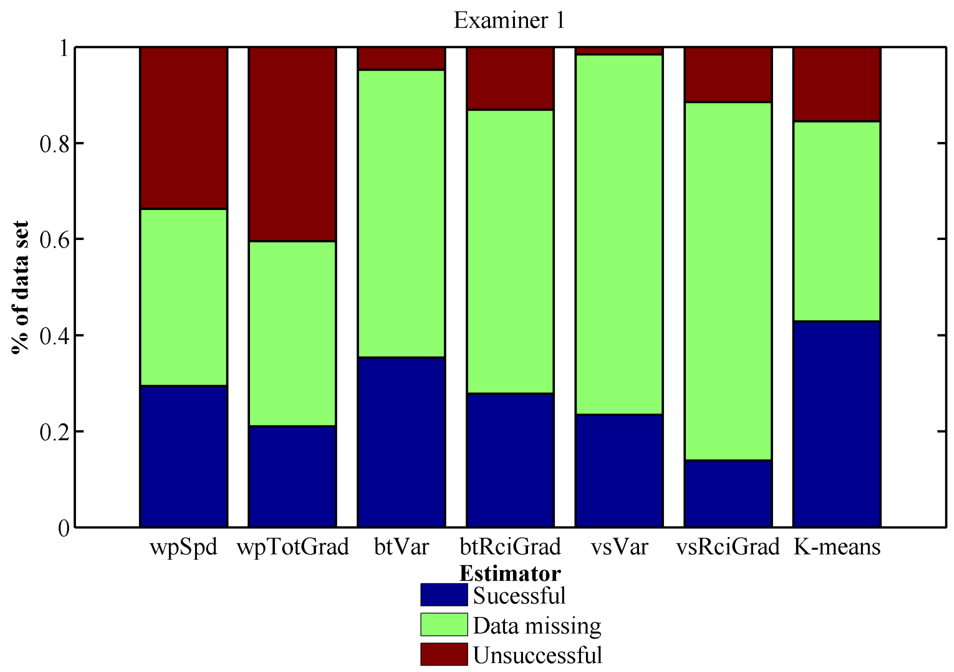


Figure 8: Results of the visual examination of the estimators. On x-axis are the estimators: wpSpd and wpTotGrad using the horizontal wind, btVar and vsVar using thresholding, btRciGrad and vsRciGrad using the wavelet transform. For each estimator, the database is divided into three categories: Estimation successful (blue), Insufficient data (yellow), Estimation failed (red).

of about 50%. Finally, the clustering method is a good settlement with a (second highest) success rate of 60% and an availability of 62% (second highest).

Contents

Introduction

The boundary layer is a region of interest for many purposes, and its vertical extent is one meaningful variable to describe it. Regarding air quality, the pollutants stay below the boundary layer top, because transport is stopped by the capping inversion. Regarding atmosphere models, boundary layer height is the limit where parametrizations may change because of turbulence or fluxes that differ in the free atmosphere. It's a critical zone for airplanes because they are both close to the ground and in a turbulent environment. On the specific context of NOAA's Atmosphere remote sensing group involvement, it will be used to calculate horizontal fluxes of aerosols and greenhouse gases over large urban areas, on the current field campaign INFlux. Assessing these fluxes allow an appraisal on the impact of human activities on their environment.

Several atmospheric variables can be used as tracker of boundary layer air (potential temperature, turbulence, aerosol concentration). Doppler lidars provide continuous information such as wind speed and direction, turbulence information and back-scattered intensity with high resolution, both in space and time. This work aims to find an algorithm able to automatically detect the boundary layer top using Doppler lidar measurements. Various methods have been implemented to achieve this purpose.

A first group of methods use peak detection to detect the boundary layer top. These methods are tuned individually to fit with a single data type. Two methods will be detailed : the Haar wavelet transform method, and the peak-based thresholding method. Haar wavelet transform changes the profile to make transitions such as boundary layer top well-defined peaks. Next, the boundary layer height is detected by the peak detection method. The peak-based thresholding method uses peaks in the original profile to calculate a minimum activity. The part of the profile above this minimum activity is considered in the boundary layer. Another type of method uses cluster analysis. This method is designed to use different data type simultaneously. The measurements points characterized by the various data are gathered by group of similarity. One of this group is defined as the boundary layer, and the border of the group defines the boundary layer top. The influence of missing data on this method will be discussed. The methods are appraised by a case study on a particular day. Their strength and weakness are demonstrated. Next, an assessment on the whole data set by visual examination will be presented.

1 | Workplace

This internship was carried out in one the lab of the National Oceanic and Atmospheric Administration (NOAA), the Earth System Research Center, located in Boulder, Colorado, USA. I was welcomed by Alan Brewer and Michael Hardesty in the Atmospheric Remote Sensing group, and more specifically in the team in charge of Doppler lidar[1] measurements.

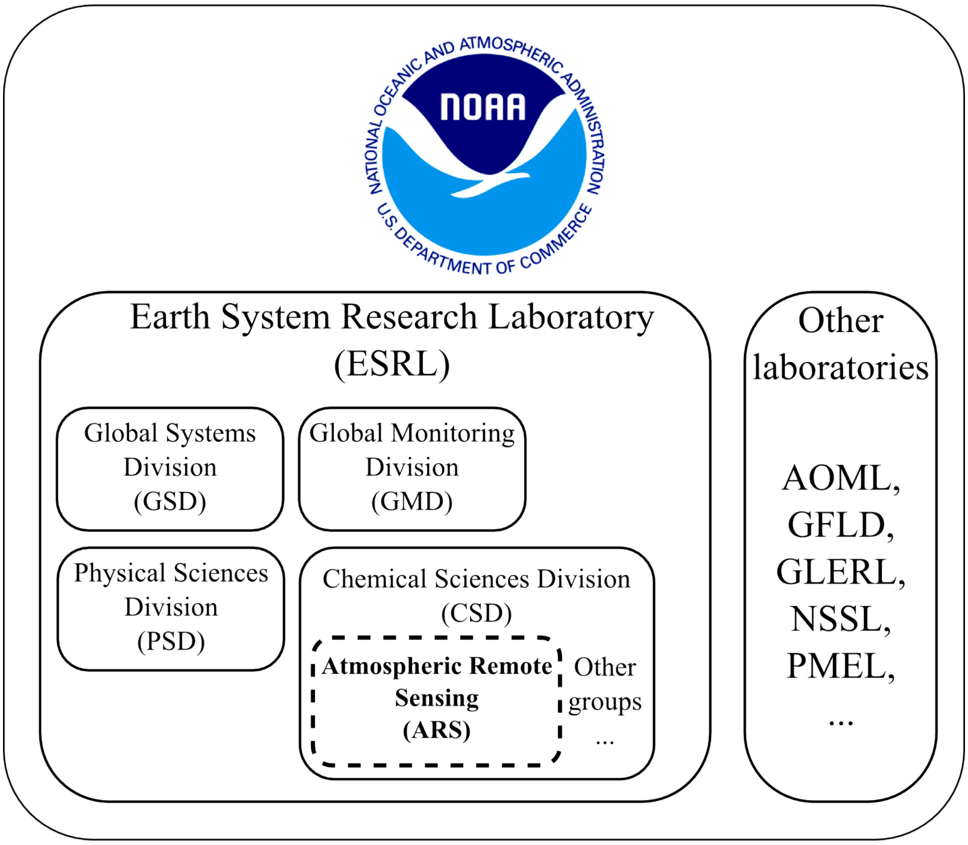

NOAA
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
U.S. DEPARTMENT OF COMMERCE
Earth System Research Laboratory (ESRL)
Global Systems Division (GSD)
Global Monitoring Division (GMD)
Physical Sciences Division (PSD)
Chemical Sciences Division (CSD)
Atmospheric Remote Sensing (ARS)
Other groups ...
Other laboratories
AOML, GFLD, GLERL, NSSL, PMEL, ...

Figure 1.1: ARS group in the context of NOAA organisation structure.

NOAA is the American federal agency in charge of national weather service. The Earth System Research Laboratory (ESRL) is one of the research laboratories of NOAA. ESRL was formed to pursue a broad and comprehensive understanding of the Earth system. This system comprises many physical, chemical and biological processes that need to be dynamically integrated to better predict their behavior over scales from local to global and periods of minutes to millennia.

Inside ESRL are four groups : CSD, PSD, GSD, GMD, represented on the organization structure of the research at NOAA, figure 1.1 (acronyms are extended in there). The Atmospheric Remote Sensing (ARS) is a subgroup of the Chemical Sciences Division (CSD). Their activity is centered on lidars. The ARS group is involved in many field campaigns over the country, and the data collected is used in various domain of research like air quality, validation of model parametrisation, wind energy, etc. This group was one of the first to use lidar for atmospheric measurement, with the lidar TEACO[2] in the 1970's, and they still keep a recognized leadership on this technology.

[1] "LIDAR" is the acronym of LIght Detection And Ranging.

[2] TEACO : Transversely Excited Atmospheric CO_2 laser.

Among the field campaign they are involved in, some are devoted to evaluate horizontal fluxes: TXFlux (Texas Flux Study) aimed to evaluate horizontal fluxes of methane downwind of large oil or gas field, INFlux (Indiana Flux Study) aims to assess horizontal fluxes of greenhouse gases and aerosols over large urban area. The horizontal fluxes occur mostly into the boundary layer, because the capping inversion stops vertical diffusion. The boundary layer top is thus a needed parameter to calculate horizontal fluxes. The automatic algorithms to evaluate boundary layer top presented here was built on the TXFlux experiment data set. It will be tested on the INFlux experiment, and the next experiment, LUMEx (Lidar Uncertainty Measurement Experiment) will allow a full assessment of these methods by comparing the results to independent and trustworthy measurement (met mast, radiosondes, others lidars).

2 | Background

2.1 Instrument and data

Data for this study are taken from a field measurement in Fort Worth, Texas, in 2013. The main purpose of the campaign (named TXFlux) was the study of horizontal fluxes of methane, downwind of large oil or gas fields. The days available are partly in early spring (from 24$^{\text{th}}$ of March 2013 to 5$^{\text{th}}$ of April 2013) and partly in autumn (from 22$^{\text{nd}}$ to 29$^{\text{th}}$ of October 2013). The data set is 21 days long and is taken under various conditions.

2.1.1 Measurements from HRDL

The data I used are provided by the lidar HRDL for High Resolution Doppler Lidar. It's a pulsed Doppler lidar, with a 2 μm wavelength. Several papers document this lidar, for example Grund [1997] or Grund et al. [2001]. Its characteristics are summarized in the table 2.1.

The wavelength defines which targets will back-scatter the light. A 2μm wavelength corresponds to infrared light, eye-safe, and Mie scattering. The targets are aerosols and small dusts in the atmosphere. The laser pulse energy is the energy sent in the atmosphere by 1 pulse. The laser pulse width is the width time for 1 pulse at half the maximum. The pulse repetition rate is the number of pulses sent in 1 second (200). The beam rate is the number of range resolved measurements in 1 second. It differs from the pulse repetition rate because 100 pulse are averaged to get 1 range-resolved measurement (the average reduces the noise). The range resolution is the minimum distance between two measurements. The minimum range depends on the pulse width and the post-processing. The maximum range depends essentially on the aerosol content. Vertically staring, HRDL can see as long as there is aerosols. The figure given here is the length of the available data set.

Parameter	Value
Wavelength (λ)	2.018 μm
Laser pulse energy	2 mJ
Laser pulse width	200 ns
Pulse repetition rate	200 Hz
Beam rate	2Hz
Range resolution	30 m
Minimum range	260 m
Maximum range	10.8 km

Table 2.1: HRDL basic characteristics (from Grund et al. [2001], updated)

The scanning strategy was split into 3 types of scans, with a pattern repeat period of 20 min. So, in the algorithms, 20 minutes is considered as the temporal resolution.

The first type of scan is conical. The lidar looks over from 0° to 360° azimuths, for 3 different elevations angles. This kind of scan is analyzed to provide vertical profiles of wind speed and direction. The velocity azimuth display (VAD) technique is applied to retrieve vertical profile from the raw data (see paragraph 2.1.2, and Browning and Wexler [1968]).

The second type of scan was suited to measure the layers very close to the ground. It's a partial vertical scan focused on the lowest layers. The lidar looks over the elevation angles from 0° to 24°, for two orthogonal azimuths. The vertical profiles are calculated by computing the statistics of horizontal velocity within horizontal slices 30 meters thick. The thickness is chosen to match the 30 meters vertical resolution imposed by vertical staring (HRDL's range resolution). It is called "bowtie" regards to its shape.

To complete the pattern, the lidar keeps staring vertically. This is the simplest scan, but it provides very useful information. All the data about vertical velocity come from this scan. In addition, the measurement volume can go very high when it is directed vertically. Hence, the vertical staring has a very high maximum range. It's limitation is the minimum range. To avoid too powerful return into the optics, the sensitivity of the lidar is reduced for a short period after the outgoing pulse. It yields to a minimum range around 260 meters.

2.1.2 Data retrieval

A lidar measures the optical power that is back-scattered from the atmosphere. According to the wavelength, the targets that sends the light back are aerosols. The lidar measures the motion of aerosols, which is assumed to be equal to the air motion. We will see now the techniques that retrieve the atmospheric information from this optical power. These techniques are specific to the type of lidar. A Doppler lidar will measure two parameters: the Doppler shift, from which we get the wind information, and backscattered intensity, form which the aerosol layering is infered.

Wind information The wind information is deduced from the Doppler shift. The movement of particles scattering the light back to the lidar yields a change in the light frequency. But this change is small (few 10 MHz) compared to an optical frequency (few 10 THz), so the heterodyne technique is applied to measure it accurately. The principle of heterodyne measurement is to mix the optical signal that comes back from the atmosphere with a reference optical signal. We mix both the emitted and received signal on the receptor. Let's say that our original signal has a frequency f_0. The electric field of the reference signal can be written as in equation 2.1.

$$S_0 = E_0 \sin(2\pi f_0 t) \tag{2.1}$$

The optical signal that comes back from the atmosphere will have a form like in equation 2.2. Where $\Delta f_{Doppler}$ is the Doppler shift. As we mentioned before, we can easily assume that $\Delta f_{Doppler} \ll f_0$

$$S_{back} = E_{back} \sin\Big(2\pi(f_0 + \Delta f_{Doppler})t\Big) \tag{2.2}$$

The heterodyne method entails adding the reference signal to the received one on the sensor. The sensor is only sensitive to the intensity, which is the averaged square of the signal :

$$I = \Big\langle (S_0 + S_{back})^2 \Big\rangle = \Big\langle S_0^2 + S_{back}^2 + 2\, S_0 S_{back} \Big\rangle$$

The cross-product in the only term of interest, and it is developed in equation 2.4. The Fourier spectra of this cross-product is bimodal, thanks to the trigonometric properties.

$$\begin{aligned} S_0 S_{back} &= E_h \sin(2\pi f_0 t) \sin\Big(2\pi(f_0 + \Delta f_{Doppler})t\Big) & (2.3)\\ &= E_h \frac{1}{2}\Big[\cos(2\pi \Delta f_{Doppler} t) - \cos\Big(2\pi(2\, f_0 + \Delta f_{Doppler})t\Big)\Big] & (2.4)\end{aligned}$$

The second cosine wave in the equation 2.4 has a frequency in $2\, f_0 + \Delta f_{Doppler} \simeq 2\, f_0$. This frequency is very high, compared to the first cosine wave, which frequency is $\Delta f_{Doppler}$. Now, with a low-pass filter, we can keep only the Doppler shift, and deduce the wind speed in the direction of the beam (v_{los} for v "line-of-sight") from the relation 2.5.

$$\Delta f_{Doppler} = \frac{2 v_{los}}{\lambda} \tag{2.5}$$

As the reader may notice, we can only measure the wind in the direction of the beam. This property is very useful if we want to measure the vertical wind. A lidar staring vertically will not be affected by the

VAD scan

- **Provides :** Horizontal wind speed profile (W) and wind direction (θ)
- **Typical minimum range :** 30 m.
- **Typical maximum range :** 3000 m.
- **Notation :** WP (WP.spd = W, WP.dir = θ).

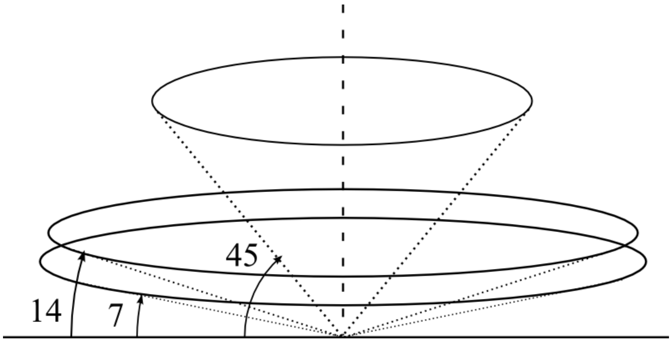


Figure 2.1: VAD scan with 3 elevations angles : 7° , 14° and 46° .

Bowtie scan

- **Provides :** Horizontal velocity variance (σ_h) and range-corrected intensity (β_{BT})
- **Typical minimum range :** 30 m.
- **Typical maximum range :** 1000 m.
- **Notation :** BT (BT.var = σ_h, BT.RCI = β_{BT})

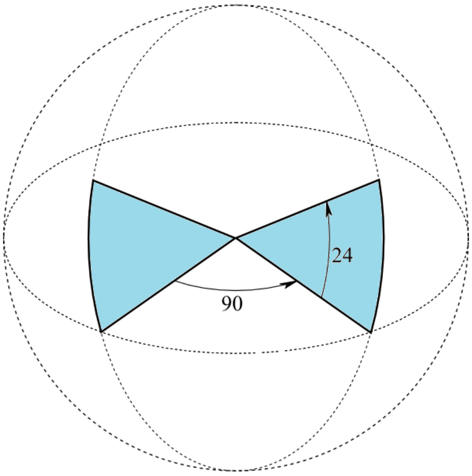


Figure 2.2: Bowtie scan from 0° to 24° elevation, and 2 orthogonal azimuths.

Vertical staring

- **Provides :** Vertical velocity variance (σ_w) and range-corrected intensity (β_{VS})
- **Typical minimum range :** 260 m.
- **Typical maximum range :** 10000 m (or top of the aerosol layer).
- **Notation :** VS (VS.var = σ_w, BT.RCI = β_{VS})

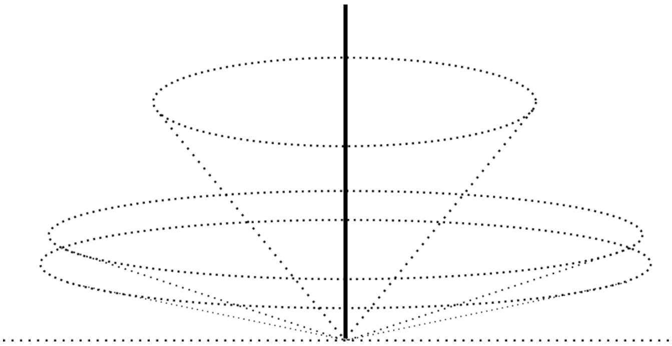
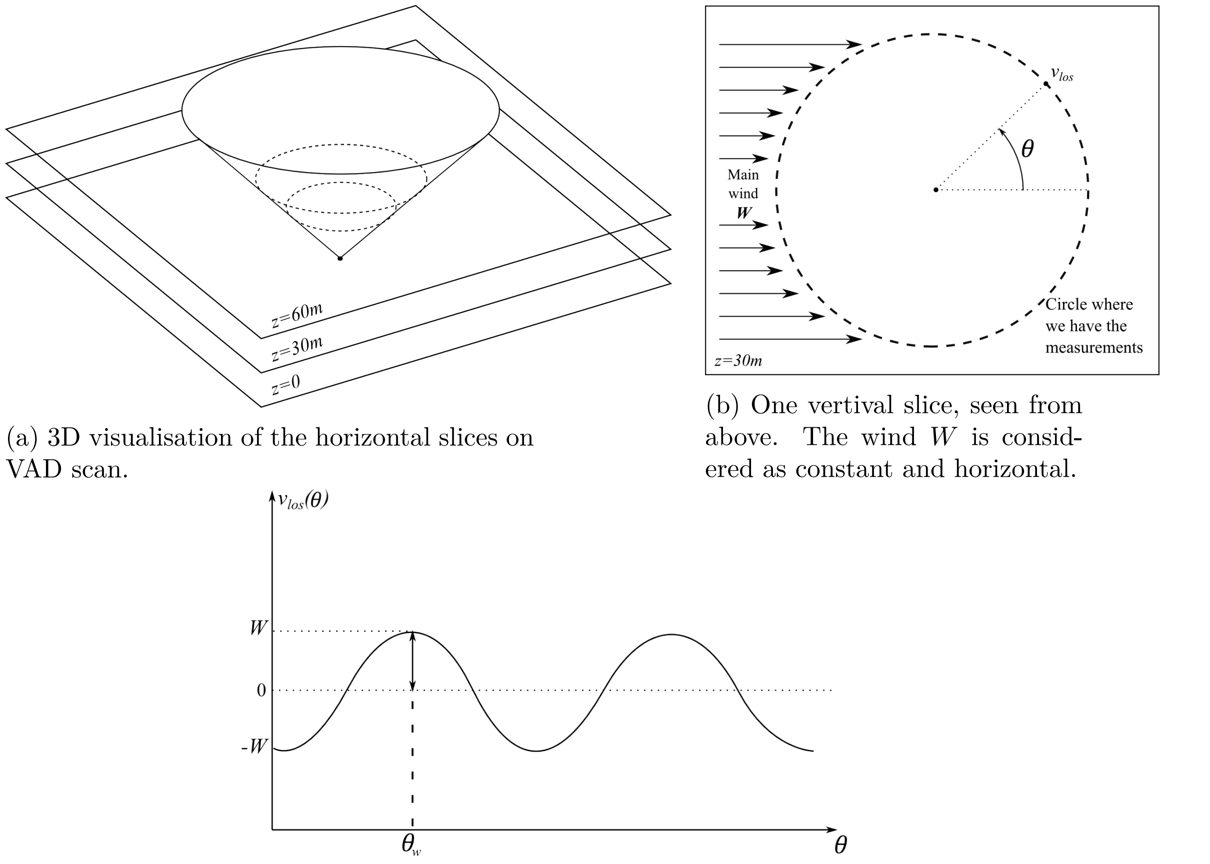


Figure 2.3: Vertical staring (superimposed onto dotted VAD scan)



(a) 3D visualisation of the horizontal slices on VAD scan.

(b) One vertival slice, seen from above. The wind W is considered as constant and horizontal.

(c) Evolution of v_{los} with respect with θ on a horizontal slice. The elevation angle is neglected.

Figure 2.4: Simplified example of data processing on VAD scan. The elevation angle is supposed to small enough to be neglected, and the wind is considered as constant and horizontal.

horizontal wind, though it is 2 orders of magnitude stronger[1]. But if we are interested in the horizontal wind, it's impossible to measure it from a vertical staring. That is the aim of velocity azimuth display (VAD) scan (see figure 2.1).

VAD is the name of the retrieval technique used to calculate vertical profiles of horizontal wind from conical scan. The vertical profile is calculated by cutting the cone in horizontal slices, as shown in figure 2.4a. In an horizontal plane, a circle of data is at the same height. For each point in the circle, the wind in the line-of-sight v_{los} is known. We consider a mean wind W in a direction θ_w (direction where it comes from). It is assumed that the wind is constant on all the horizontal plane (to keep it simple). v_{los} will be equal to the mean wind only when the beam and the mean wind are in the same line. Rigorously, we have to take the elevation angle into account. This implies the use of spherical coordinates, in a Galilean referential. For the sake of simplicity, we assume here that the elevation angle is little enough to neglect its influence. A rigorous formulation of the problem is provided by Browning and Wexler [1968]. v_{los} and W are in the line when $\theta = \theta_w$ and $\theta = \theta_w + \pi$. In the first case, the wind goes toward the lidar, so $v_{los}(\theta_w) \simeq W$ because of the small order of magnitude of the vertical wind, compared to the horizontal one. In the second case, the wind goes outward the lidar, so $v_{los}(\theta_w + \pi) \simeq -W$. When the beam is perpendicular to the wind, $v_{los}(\theta_w \pm \frac{\pi}{2}) \simeq 0$. To summarize, $v_{los} = -\overrightarrow{b} \cdot \overrightarrow{W}$, where $\overrightarrow{b}$ is the unit vector aligned to the beam. Plotting v_{los} as a function of θ (see figure 2.4c), leads to a sinusoidal curve where the magnitude is proportional to the mean wind, and the phase of the maximum indicates the direction where it comes from. The interested reader will find more details in Browning and Wexler [1968].

[1] Considering the orders of magnitude for horizontal wind $W_h \sim 1\text{m} \cdot \text{s}^{-1}$ and for vertical wind $w \sim 1\text{cm} \cdot \text{s}^{-1}$ (without convection)

Aerosol content The light emitted in the atmosphere comes back to the lidar because it's reflected onto aerosols. The relationship between the reflected light and the aerosol content is given by the lidar equation.

$$P_r(Z) = P_{pulse} T_{inst} T_{atm}(Z) \beta(Z) \frac{A}{Z^2} \tag{2.6}$$

where

- $P_r(Z)$ is the optical power received from the layer at the distance Z.
- P_{pulse} is the optical power of the initial pulse.
- T_{inst} is the total (round-trip) transmissivity of the instrument (constant).
- $T_{atm}(Z)$ is the total (round-trip) transmissivity of the atmosphere.
- $\beta(Z)$ is the scattering coefficient of the layer at Z.
- A is the area of the optical collector.

T_{atm} depends on Z according to the Bouguer-Lambert's law :

$$T_{atm}(Z) = \int_0^Z \alpha(z) dz \tag{2.7}$$

where $\alpha(z)$ is the absorption of the layer z.

What is important is this equation is that the number of photons coming back decreases as a function of the range. The usual quantity used to measure the aerosol content in the atmosphere is the range-corrected intensity (RCI). It is the intensity coming back ($P_r(z)$) normalized by the range function ρ. For direct detection lidars, the range function is $\rho_0 \propto \frac{1}{z^2}$, this leads to a range corrected intensity $\frac{P_r(z)}{\rho_0} \propto z^2 P_r(z)$. For HRDL, a complexity is added because of the heterodyne detection (see previous paragraph). Indeed, since we have both a reference signal and atmospheric signal on the detector, we don't measure $P_r(z)$ properly. The range function is modulated by the heterodyne efficiency ξ. This efficiency is low for close targets, and increases as a z^2 function until a bound called the Rayleigh range. On the figure 2.5, we can see these range functions and the heterodyne efficiency. The blue curve is the range function for direct detection ρ_0, that varies as $\frac{1}{z^2}$. The heterodyne efficiency is the red curve (the scale is arbitrary). It increases asymptotically as a z^2 function and raises a maximum after the Rayleigh range. The last curve (black one) is the HRDL range function ρ_H. It is the products of the two previous.

In the end, the range-corrected intensity for HRDL, that we will note β (from "backscatter") is given by the next formula :

$$\beta = \frac{P_r(z)}{\rho_H} = \frac{P_r(z)}{\rho_0 \, \xi}$$

In practice, the range function of HRDL is measured by low elevation angle horizontal scans where we assume the aerosol content as constant. The main thing to remember, is that the range-corrected intensity is proportional to aerosol content most of the time (it is not when the beam hits clouds, rain or hard targets, for example).

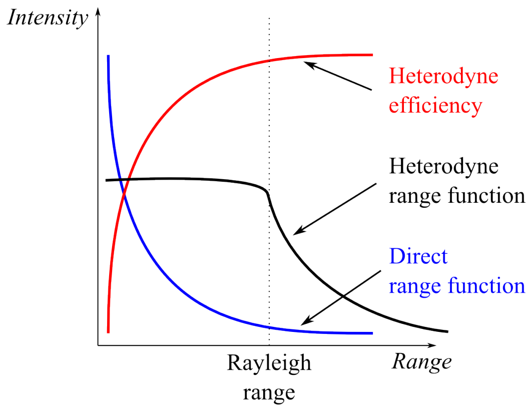


Figure 2.5: Theoretical range function for HRDL ρ_H (black thick line). In blue, the range function for direct detection ρ_0 (decrease as a function of $\frac{1}{z^2}$). In red, the heterodyne efficiency ξ (asymptotic increase as a function of z^2). The figure is not to scale.

Type of scan	Vertical staring (VS)		Bowtie (BT)		Wind profile (WP)	
Data	Range-corrected intensity	Vertical velocity variance	Range-corrected intensity	Horizontal velocity variance	Wind speed	Wind direction
Notation	β_{VS}	σ_w	β_{BT}	σ_h	W	θ

Table 2.2: Summary of the different lidar data used and their notations.

2.2 Boundary layer

2.2.1 Diurnal cycle

A widely used definition of boundary layer is given by Stull [1988] :

> The boundary layer is the part of the troposphere that is directly influenced by the presence of Earth's surface and responds to surface forcings with a timescale of about an hour or less.

To reiterate, the purpose of the field campaign TXFlux (and INFlux, the current one) is too assess the horizontal fluxes of aerosols and greenhouse gases like methane. As a consequence, if the definition of the boundary layer is subject to uncertainty, the layer where the aerosols stay after being released from the ground will be chosen.

Boundary layer (shortened in BL thereafter) is well known to present different characteristics according to meteorological conditions ([Stull, 1988], [Harvey et al., 2013], ...). As the definition says, the BL is widely influenced by Earth's surface, and thus the BL have a stronger variability than the layers above. The diurnal cycle is one of most important cause of variability. The figure 2.6 sums up the different behaviors of BL along the day.

During daytime, the sun brings energy into the earth system. The solar heating doesn't warm the boundary layer directly because of the transparency of the air. The air is heated by the contact with the ground (conduction and convection). As a consequence, the warmest layers of air are under cooler ones. Vertical movements are very likely to appear in such conditions, creating eddies at many range of scales. From these eddies results a turbulent transport that tends to mix heat, moisture and aerosols. That's why daytime BL is also called "mixed layer" (ML) or "convective boundary layer" (CBL) : the layer where turbulence is convectively driven and forms a homogeneous volume of air.

During nighttime, the Earth's surface has a cooling effect because of radiation balance [Stull, 1988]. This cooling effect creates a stable stratification near the ground. It tends to extend vertically during the night because the higher layers have no longer energy source (the ground). Stability is the highest at the ground level and decreases with altitude. Since the conditions are stable, vertical movements die away, and the transport is more likely horizontal. The horizontal wind speed also tends to decrease at the ground level. But a few hundred meters above, an acceleration of the wind is likely to occur. This phenomenon is called a nocturnal (low-level) jet or simply low-level jet (LLJ). Above the stable layer there are usually less stable layers where residuals of the former mixed layer remain (homogeneous plumes). Hence the name of "residual layer" (RL). But the RL is not properly part of the boundary layer, according to the definition, because it is disconnected from the ground by the stable layer. However, the RL is an important component of the nocturnal behavior of BL. Complex and heterogeneous structures may appear in stable conditions. With such a fuzzy structure, finding the limit of the stable boundary layer (SBL) is not an easy task.

2.2.2 Tracking the boundary layer air

As previously described, though the BL definition doesn't change, the boundary layer air is characterized differently. On figure 2.7 is shown two example of profiles. They are range-corrected intensity profile and velocity variance profile. One is taken at night (green profiles, 2.7a) and the other is taken at day (red profiles,2.7b). The ranges are different between day and night because the data come from different scans[2].

During daytime (figure 2.7b), the boundary layer is characterized by high values of range-corrected intensity and velocity variance. It corresponds to a high aerosol content and high turbulence. Indeed, the ground is the first source of aerosols, and also create turbulence by friction. Aerosols and turbulence are correlated because the turbulence is the vertical transport mechanism of the aerosols. The vertical velocity variance is more relevant than horizontal variance because of convection. The limit with the free troposphere is a sharp gradient in the profile.

During nighttime (figure 2.7a), there is still the highest values in a shallow layer close to the ground. But the transition is not obvious, and it's debatable whether this shallow layer shall be called a boundary layer. As mentioned in the previous section, we are interested in the layer where the emitted aerosols remain. According to the range-corrected intensity profile, this shallow layer is the boundary layer to be detected. The shape to detect is still a step in both profile, but at night, it is less sharp than during the day. A look at the values of velocity variance indicate us a poor turbulence. The horizontal variance of the wind is here a better signal because the vertical activity is close to zero. The wind information could be a good indicator. Low-level jet, for example, define usually the top of the boundary layer [Pichugina and Banta, 2010]. The shear can also be used as an indicator of the boundary layer top [Tucker et al., 2009]. But the shortcoming is that they don't occur under every condition.

[2] At night BT is preferable, while at day it is VS. More information about their specificity is given in section 2.1.

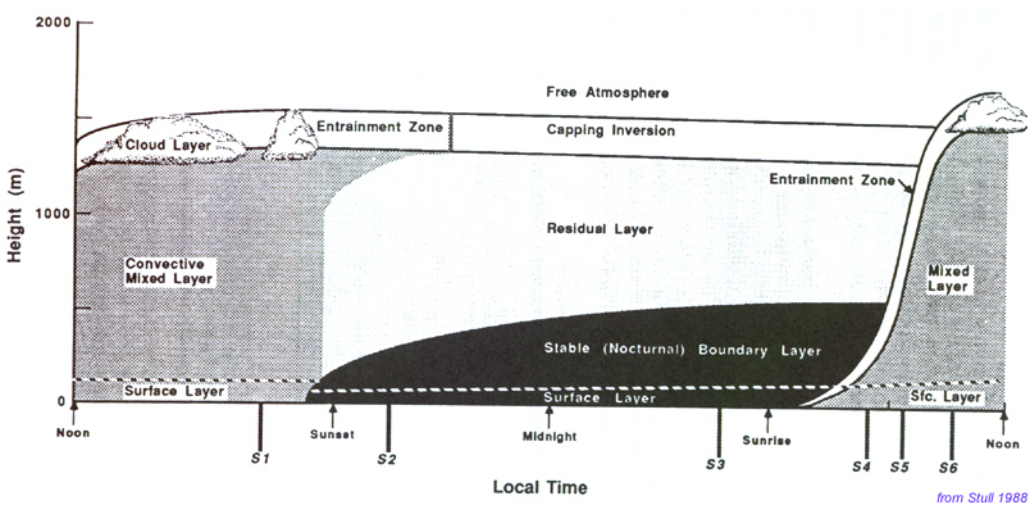
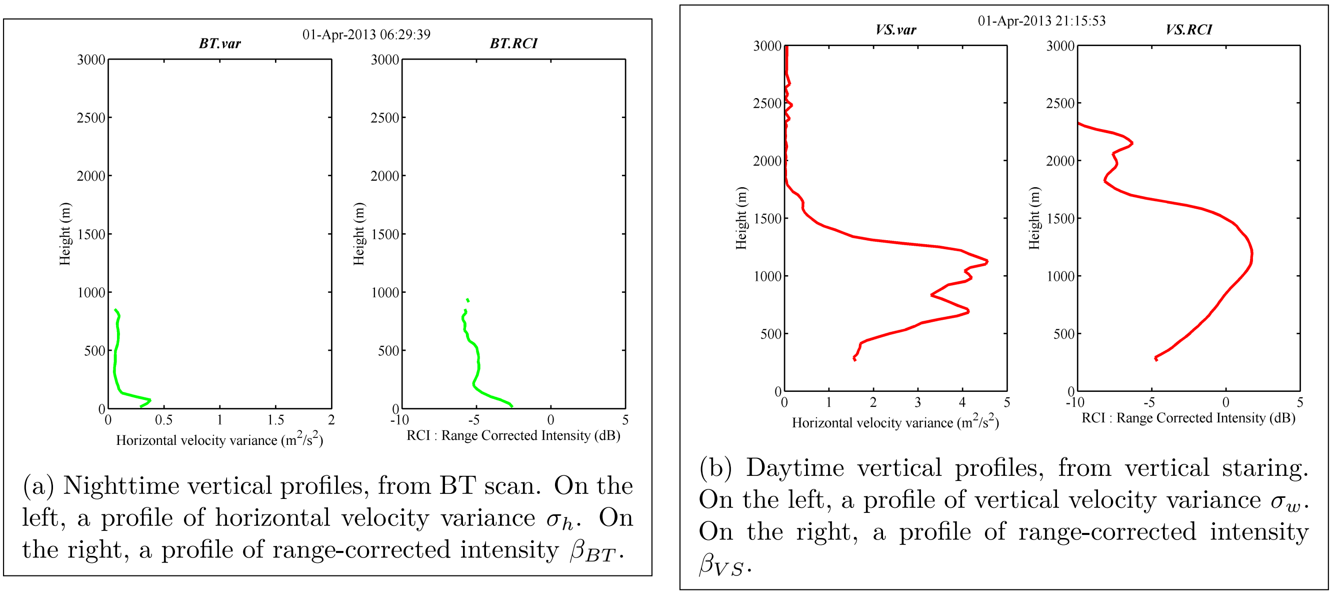


Figure 2.6: Schematic of the boundary layer during a diurnal cycle : mixed layer during daytime, stable and more complex structure over night. (from Stull [1988])



(a) Nighttime vertical profiles, from BT scan. On the left, a profile of horizontal velocity variance σ_h. On the right, a profile of range-corrected intensity β_{BT}.

(b) Daytime vertical profiles, from vertical staring. On the left, a profile of vertical velocity variance σ_w. On the right, a profile of range-corrected intensity β_{VS}.

Figure 2.7: Example of profile during nighttime and daytime.

3 | Methods developed

3.1 Peak detection methods

This section will presents two methods to evaluate BLH based on peak detection in the relevant signal. Signal here means processed signal (a vertical profile), not the raw signal from the lidar. For the first method, this relevant signal is a gradient-like transformation. The second method looks for peaks in the profile itself. The third subsection presents a processing of output data that look for continuity in the peaks.

Definition 1 (*Peak*) *Let's consider a discrete profile $f(z^i)$ where z^i are the points of measurements. A* peak *in this profile, is a point of measurement p for which $f(z^p)$ is higher than its neighbors*

$$f(z^p) > f(z^{p+1}) \wedge f(z^p) > f(z^{p-1})$$

A pre-filtering is applied by keeping only one peak on a 90 meters moving window. If a conflict occurs, we keep the highest peak in the window.

3.1.1 Haar wavelet transform method

As presented in section 2.2.2, the boundary layer top may look like a step in vertical profiles (like in RCI profile, fig. 3.1a). To detect this step, several techniques can be used. A family of techniques transform the original profile to transform the step into a peak. Then, we detected the step by detecting the peak in the transformed signal. This section will introduce the transformation applied before the peak detection. The literature has many examples on this topic because is a way to process back-scattered intensity, which is provided by most lidars (both direct and heterodyne detection).

First of all, the level of the step can be defined as the level where the profile drops below a given threshold. For convective and well defined BL, it is a very simple way to have a continuous BLH measurement, according to Melfi et al. [1985]. Nonetheless, the choice of a fixed value for the threshold fails to adapt to the boundary layer state.

Another way is to find the extremum in the gradient profile. This method have the advantage to fit exactly with the intuitive definition of a step. The step will appear as a peak in the gradient profile. Many authors used it directly (for example Hayden et al. [1997]) or with some variances. Menut et al. [1999] used the height of zeroing the second derivative (inflection point). Senff et al. [1996] used the derivative of the logarithm of the back-scattered intensity along the height. The limitation of these methods is the gradient calculation, which is sensitive to noise, and can be poisoned by small scale structure [Gamage and Hagelberg, 1993].

Inspired from the derivative methods, but different enough to fix some shortcomings, are the variance methods. Instead of computing the vertical gradient of the signal, Menut et al. [1999], for example, compute the variance of the signal and define the boundary layer top where the variance is maximum. Another variance-based method doesn't compute the vertical variance of the signal but a covariance : the covariance between the step-like profile and a Haar wavelet. It's the techniques chosen for this work. The table 3.1 summarizes the step detection techniques evoked before.

The Haar wavelet transform will be used here as a method to calculate an "improved gradient". The wavelet covariance transform (WCT) as defined by Gamage and Hagelberg [1993] is a way to detect step change in a signal. It's based upon a step function: the Haar function ψ_H. This Haar function depends

Family of method	**Definition of BLH**	**Advantages**	**Drawbacks**
Direct threshold	first $z\ /\quad S(z) < S_0$	Simple and fast	Choice of the threshold
First derivative	$arg\min\left(\frac{\partial S}{\partial z}(z)\right)$	High efficiency	Sensitive to noise
Second derivative	$z\ /\quad \frac{\partial^2 S}{\partial^2 z}(z) = 0$		
First derivative of $\log S$	$arg\min\left(\frac{\partial \log S}{\partial z}(z)\right)$		
Vertical variance	$arg\max\Big((S * S)(z)\Big)$	High efficiency, poorly affected by noise	Choice of the window
Haar wavelet transform	$arg\max\Big((\psi_H * S)(z)\Big)$		Choice of the dilatation

Table 3.1: Summary of the different method to detect a step in a profile $S(z)$ where S is the signal measured by the lidar (mostly backscattered intensity), S_0 is the threshold value and ψ_H is the Haar wavelet.

on 2 parameters: a its dilatation, and b its translation. It is defined for an altitude z as follow :

$$\psi_H\left(\frac{z-b}{a}\right) = \begin{cases} -1, & \text{if} \quad b - \frac{a}{2} \leqslant z \leqslant b \\ 1, & \text{if} \quad b \leqslant z \leqslant b + \frac{a}{2} \\ 0, & \quad \text{elsewhere} \end{cases} \tag{3.1}$$

An example of Haar wavelet is drawn on figure 3.1b. The wavelet covariance transform described by Brooks [2003] is given by

$$W_f(a,b) = \frac{1}{a}\int_{-\infty}^{\infty} f(z)\psi_H\left(\frac{z-b}{a}\right) dz \tag{3.2}$$

where $f(z)$ is our vertical profile.

We can see this transformation as a convolution : $W_f(a,b) = (f * \psi_H)(b)$. The quantity $W_f(a,b)$ is so a kind of measurement of the similarity between f and $\psi_{H(a,b)}$. As a consequence, if we plot the points $\{W_f(a,b),\ a\,\text{fixed},\ b \in [0, Z_{max}]\}$, we will see a vertical profile of WCT[1] as the example shown on figure 3.1c. The WCT shows high value in the transition zone, where is the step, and small values elsewhere. Compared to the direct gradient on 3.1d, we can see 3 noticeable differences. First, is the order of magnitude. Values on the x-axis for the gradient range from -0.06 dB·m^{-1} to 0.02 dB·m^{-1}, while the WCT ranges from -1 to 2. It's 2 order of magnitude above. Here is one reason why WCT is less sensitive to noise. Second, the top of the profile (above 2500 m). It is more variable in the gradient profile than in the WCT profile. This is one argument more to prefer the WCT. And third, the WCT has positive peaks where the gradient has negative ones. This is only because the Haar wavelet starts with negative values.

Nevertheless, the WCT the choice of the dilatation a is an issue. We can see on the figure 3.2 the wavelet covariance transform of the same profile as figure 3.1 (fig. 3.2a), but computed for different dilatations a (fig. 3.2b). The dilatations used for this computation range from 60 m (the minimum dilatation : twice the vertical resolution of the lidar) to 1200 m. For large dilatations, the WCT profile is very smooth and is very likely to miss some transitions in the original profile. In addition, the maxima are not sharp enough to make an efficient criterion and the layering given by the WCT is too coarse to fit with the reality. For small dilatations, the WCT is more turbulent and sensitive to noise. It can be as noisy as the gradient in the top of the profile, for example (a = 60m). The magnitude of the WCT is also low, which makes more difficult to distinguish the peaks due to atmospheric layering among the peaks due to the noise. According to Brooks [2003], the best dilation value is close to the depth of the

[1] WCT = Wavelet Covariance Transform

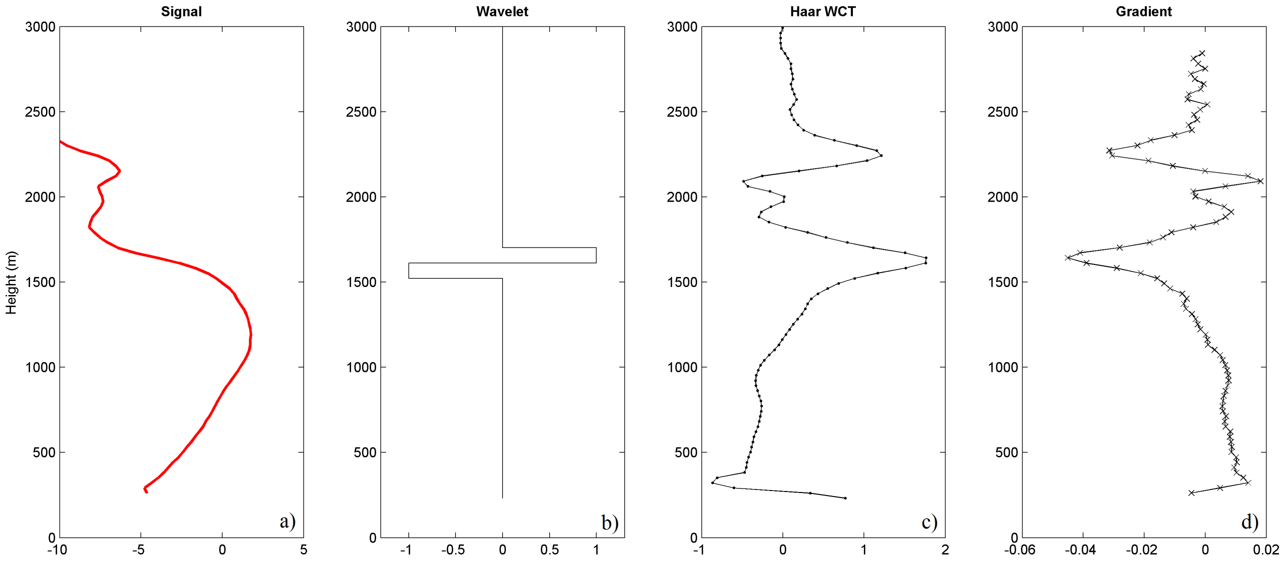


Figure 3.1: Example of Haar WCT applied on a backscatter profile and comparison to the gradient. (a) Original profile (RCI, 01-Apr-2013 21:15 UTC). (b) The Haar wavelet used for the calculus. (c) The WCT $W_f(a, b)$. (d) The direct gradient. Notice the very small values of the gradient compared to the WCT and the noise above 2500 m.

entrainment zone, which is *a priori* unknown. Our choice is a dilatation $a = 180$ m, based on the visual examination of the best fit with height of the maximum WCT and the height of the step.

Further information can be found in [Gamage and Hagelberg, 1993] about the mathematical aspect of this application of wavelet covariance transform, in [Brooks, 2003] about the influence of the dilatation, in [Cohn and Angevine, 2000] and about the application of this method for boundary layer height measurement.

In summary, the Haar wavelet transform is a method of choice to detect steps in a signal, but other techniques exist. A very brief overview of these techniques is given in the table 3.1. After the formal definitions of the Haar wavelet and the wavelet covariance transform, an example of WCT on a typical profile was provided, and compared to the direct gradient of this profile. We can see on the figure 3.1 that the WCT is less sensitive to the noise than the direct gradient. But this result is tempered by a discussion on the choice for the dilatation value. The WCT is applied on the original profile, and then the BLH is detected as a peak in the transformed profile.

3.1.2 Peak-based thresholding

Some profile may present some small-scale, but steep, gradient. These gradients can poison a BLH detection by the Haar wavelet method, even though it is less sensitive than the classic gradient. Vertical velocity variance profiles are good examples of profiles with sharp gradients between the peaks (see fig. 3.3). For this kind of profile, it is of interest to consider the absolute value, and not only the gradient, that could be noisy. That justifies an approach by thresholds. This second method is still based on peak detection, as well as Haar wavelet transform method, but here we look for peak directly in the profile. The main work in this method is the dynamic setting of the threshold, that will be described below.

Let's consider a profile of vertical velocity variance ${\sigma_w}^1, \dots, {\sigma_w}^n$. The height of interest is where this profile is no longer turbulent. This is characterized by a "low" velocity variance. But the notion of "low" depends on the meteorological conditions, that is to say, on the profile itself. For example, during a very convective day, the velocity variance can reach 2 $\mathrm{m}^2 \cdot \mathrm{s}^{-2}$, while it barely reaches 0.1 $\mathrm{m}^2 \cdot \mathrm{s}^{-2}$ under stable conditions. We assume that the profile provides enough information to deal with this variability.

The detection of the end of the turbulence will be explained on the example shown in figure 3.3. On the figure 3.3 is a vertical velocity variance profile (red solid line). Peaks, as defined in equation 3.3 (local

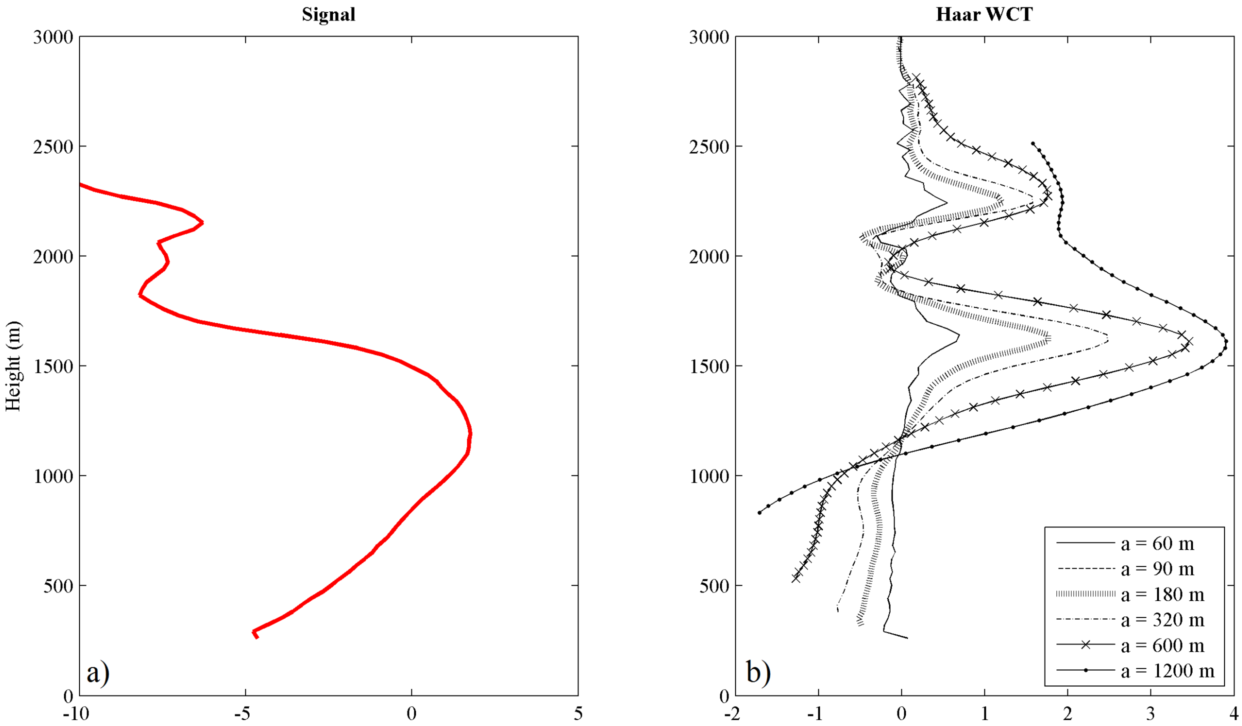


Figure 3.2: Influence of the dilatation a on the WCT.
(a) Original signal (RCI, 01-Apr-2013 21:15 UTC). (b) WCT of the signal for several dilatations. The dilatations tested are $a = 60$ m (solid line), $a = 90$ m (dashed line, unfortunately superimposed to the solid line), $a = 180$ m (thick dotted line), $a = 320$ m (dash-dotted line), $a = 600$ m (solid-crossed line) and $a = 1200$ m (solid-dotted line).

maximum), are detected in this profile (black triangles). To filter the small scale effects, only the peaks above a given threshold value σ_w^0 (dotted line) are kept. The value of this threshold is chosen low enough to ensure that no peak in the boundary layer are filtered. The threshold σ_w^0 has a double use. It's a low bound for the peak values, but it also defines the *background level*. The background level σ_w^B is the median velocity variance of the part of the profile below σ_w^0 :

$$\sigma_w^B \Big/ \quad \left|\left\{{\sigma_w}^i < \sigma_w^B,\ 1 \leqslant i \leqslant n\right\}\right| = \frac{1}{2}\left|\left\{{\sigma_w}^i < \sigma_w^0,\ 1 \leqslant i \leqslant n\right\}\right|$$

where $|\{\cdot\}|$ is the number of elements in the ensemble $\{\cdot\}$. σ_w^B can be interpreted as the background activity of the profile. An immediate property of σ_w^B is $\sigma_w^B \leqslant \sigma_w^0$. If the profile is turbulent, σ_w^B will be close to σ_w^0. The background activity is important. If the profile is stable, σ_w^B will be very small, close to 0. In the end, σ_w^B is a modulation of the given threshold σ_w^0 by the profile characteristics. The background level is only function of the profile, peaks are not used yet.

The indexes where the peak are located are named $p_1, \ldots, p_m$, with m the number of thresholded peaks above σ_w^0.

$$\{p_1, \ldots, p_m\} = \left\{i \ / \ {\sigma_w}^i > {\sigma_w}^{i+1}, {\sigma_w}^{i-1}, \sigma_w^0\right\} \tag{3.3}$$

It is important to know if these peaks are connected to the ground or not. A peak is interesting if the part of the profile below this peak maintains "high values" all the way to the ground. What "high values" means depends on two things: the profile (the "high values" are higher for a turbulent profile) and the peak value (the "high values" are higher for a higher peak). This yields to calculate a peak-based threshold τ_p for each peak found.

$$\tau_p = \min\left(0.9\,\sigma_w^B + 0.1\,{\sigma_w}^p\ ,\ \sigma_w^0\right)$$

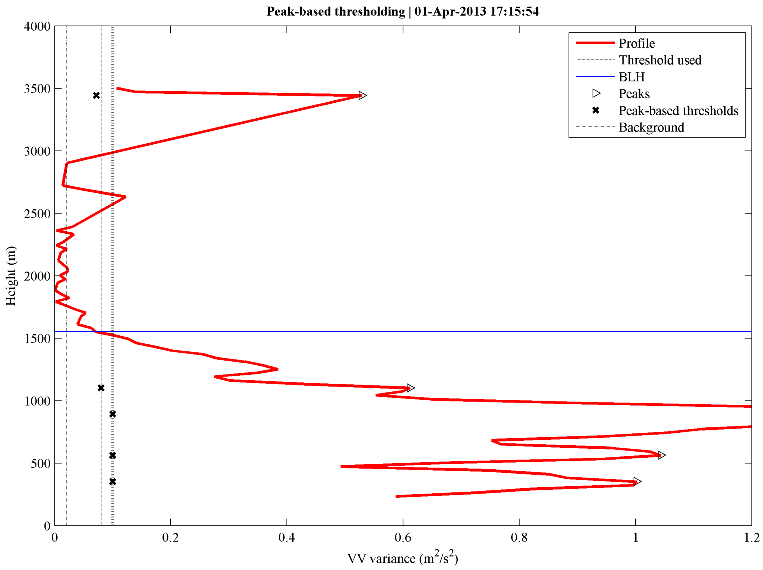


Figure 3.3: Example of peak-based thresholding.
Profile of vertical velocity variance from the 01-Apr-2013, 17:15 UTC (red solid line). The interesting peaks (triangles) are above σ_w^0 (dotted line), and for each peak p we define a peak-based threshold τ_p (crosses) depending on the peak value and the background (dash-dotted line). The highest peak connected to the ground p_{CtG} is selected, and the BLH (blue solid line) is defined as the highest point above the threshold based on this peak $\tau_{p_{CtG}}$.

This formula is similar to a weighted average of σ_w^B (background level) and σ_w^p (value of the peak p). The coefficient of 10% of the peak value σ_w^p is a setting tuned by visual examination of the result. τ_p is bounded below σ_w^0 because peaks can reach really high values that make them disconnected to the ground, though a human eye would have said that. On figure 3.3, the different τ_p are shown by black crosses.

Then the criterion for a peak p to be connected to the ground is to maintain values higher than τ_p for all the part of profile below p. As an example, on figure 3.3, the highest peak (p_m) around 3500 meters, is not connected to the ground, though it is above σ_w^0. We can see τ_{p_m}, the black cross around [0.1 $\text{m}^2 \cdot \text{s}^{-2}$, 3500 m], and the profile has values fairly below τ_{p_m} around 2800 meters an 2000 meters. There is a non-turbulent layer between this peak and the ground. It is not connected to the ground. The four lower peaks are connected to the ground. Looking for the top of the boundary layer, only the highest peak connected to the ground p_{CtG}[2] will be kept. Its threshold $\tau_{p_{CtG}}$ is chosen as the minimum velocity variance to be considered in the boundary layer. In the end, the BLH is defined as the highest point connected to the ground in the profile, which means

$$Z_{CtG} = \max\left\{z \;\; / \;\; \sigma_w(z) > \tau_{p_{CtG}}\right\}$$

On the figure 3.3 the resulting BLH is the horizontal blue line. It is observed that this method catches the good transition, though the associated gradient is not obviously steeper than the gradient between two peaks, for example.

To conclude with the peak-based thresholding, it's a method suitable for profiles that naturally present peaks. The peak values give an information about the local activity of the profile (high value in the profile). The main issue is to give a proper value to the threshold that track the limit of the boundary layer. An automatic calculation of a relevant threshold value has been implemented. The peak value is compared to a background activity to set the good threshold value. It is tuned to detect the level where there is no longer activity.

[2] The index "*CtG*" is for "Connected to the Ground". It will be used again.

3.1.3 Continuity test

The aim of the program is to run continuously. The lidar has high spatial resolution and, depending on scanning pattern, high time resolution. The pattern repeat period is 20 min in this campaign. To exploit this high temporal resolution, a piece of program is built to run on a given set of profiles (not on a single profile) and use the information of the neighboring profiles to correct the BLH estimation. This piece of program is called the "continuity test".

The idea is to replicate the human eye behavior with the program. The human eye tracks transitions with a continuous shape. For example, in RCI profile, peaks in the Haar WCT will be perceived. Among them, a strategy is need to select the peak that matches the BLH. The strongest peak (highest value in the WCT profile) could be chosen, the lowest peak (peak the nearest to the ground) as well. The purpose of the continuity test is to avoid isolated peak, even though there are legitimate.

For a given profile, few peaks can be given a method. For each peak, we look for neighboring peaks (in time, in last and next profiles). If we find any neighbor, we link the peak to its neighbors by a *thread*.

Definition 2 (*Thread, spot*) *In the continuity test, a* thread *is a link between several* spots*. A* spot *is any point of interest in a profile (peak in gradient field, transition point in cluster map) likely to be the BLH.*

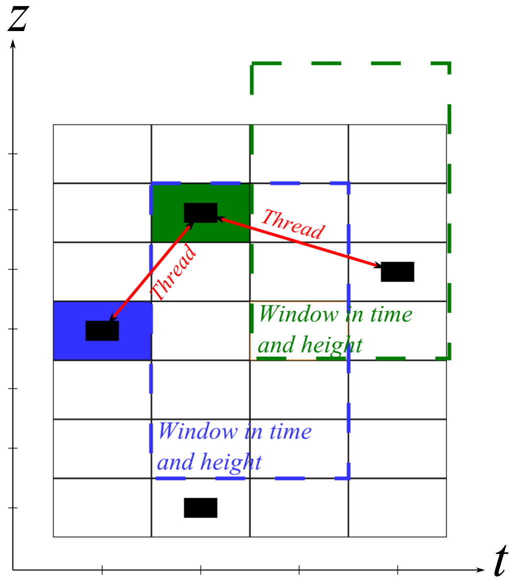


Figure 3.4: Spots linked by threads in the continuity test.
In y-axis, the height, one cell per 30 m (vertical resolution). In x-axis, the time, one cell per profile. The black-filled rectangles are spots. For two spots are shown the window in height and time where the continuity looks for any neighbor. If there is any spot in the window, the 2 spots are linked by a thread.

The figure 3.4 gives a theoretical example of a *thread*, linking three *spots*. "Windows" (dashed colored rectangles) are also visible on the figure. A window is related to one spot. For example the blue spot (black rectangle, in the blue cell of the grid), on coordinates (z_b, t_b), the window $([z_b - \Delta z, z_b + \Delta z], [t_b - \Delta t, t_b + \Delta t])$ is the blue dashed rectangle. The window defines the neighborhood where we are looking for other spots. For the blue window, (of size $\Delta z = 2$, $\Delta t = 2$), there is another spot in it: the green one. The green spot and the blue spot are then linked by a thread. Conversely, the low spot below the green one isn't in the window. Continuity test depends on 2 parameters Δz and Δt, which are the size of the window (same for every spot).

We consider now the green spot, with its window. There is another spot in the green window, thus the green spot is linked to the new spot. Since the green spot is already linked to the blue spot by a thread, this thread is extended until the new spot. As a consequence, the blue spot and the new spot are linked by a thread, though they are too far to be direct neighbors (the relationship "is linked by a thread" is transitive).

By repeating this process for every spots, a map of thread is created instead of a map of spots, which is more interesting because of the information about continuity. Excepted during transition periods, the boundary layer top moves quite smoothly, which justifies the continuity approach.

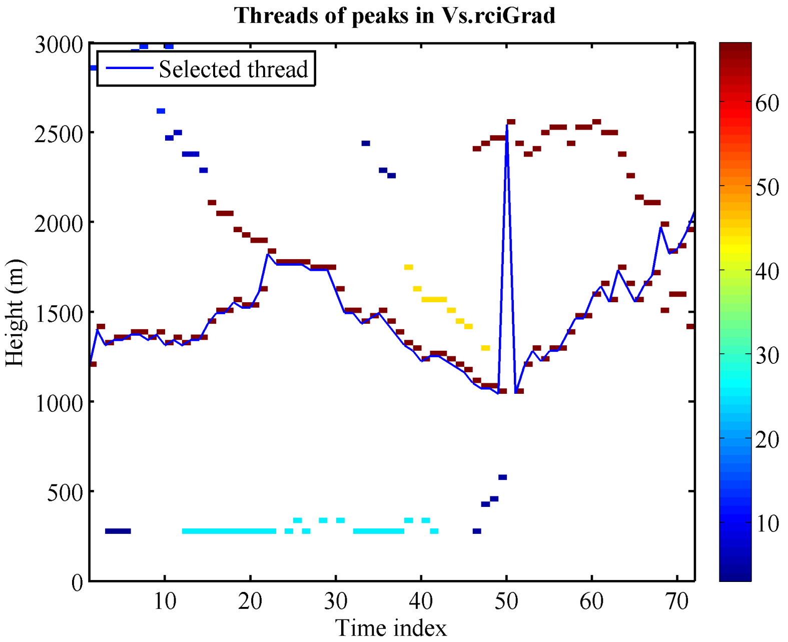


Figure 3.5: Example of thread map on a real day (01-Apr-2013). x-axis: time ; y-axis: height. The spots are peaks in the β_{VS} Haar wavelet gradient profile. The spots in the same thread are in the same color. The color bar scales the thread length, and the select thread (blue solid line) is the longest thread.

The figure 3.5 shows an example of thread map[3]. The spots observable on the map are peaks from a β_{VS} Haar wavelet gradient profile. The spots in the same thread are in the same color. The color bar scales the thread length. The blue line follows the selected thread (longest one).

As it's observed on the figure 3.5, choosing a thread could be an issue. On this figure, we choose the longest thread, to follow the maximum of continuity. We can see that during the night, the selected thread corresponds more to a residual layer than a shallow stable boundary layer. This is not a huge drawback since we know the vertical staring to be inappropriate for shallow BL. But in the morning, the selected thread goes suddenly high, while there is another thread very low, catching the early rising of the BL. The spots tracking the rising are too separated in height to be linked by any thread. As a consequence, the blue line avoids them. During transition period like the morning, the continuity test doesn't provide a good answer. The benefit of the continuity test is to filter isolated spots. On the figure 3.5, the thread map is not poisoned by unsignificant peaks.

To conclude with the continuity test, this is a tool that aims to improve the re-analyzed results by bringing an information about continuity. The continuity is followed by linking the neighboring spots with threads. The window in which we are looking for neighbor has fixed dimensions, Δz and Δt, which are the two settings of the continuity test.

[3] The time unit here is the time index. The index 1 is a midnight the 01-Apr-2013 (UTC), and the last (72) is at 23:40 UTC, with one increment every 20 min. That is to say, the night corresponds to low indexes, the sunrise occurs between 40 and 50.

Estimator	VS.RCIgrad	VS.var	BT.RCIgrad	BT.var	WP.spd	WP.totGrad
Notation	$Z(\beta_{VS})$	$Z(\sigma_w)$	$Z(\beta_{BT})$	$Z(\sigma_h)$	$Z(W)$	$Z(W,\theta)$
Type of estimation	Haar wavelet transform	Peak-based thresholding	Haar wavelet transform	Peak-based thresholding	Peak in the profile	Haar wavelet transform
Data used	β_{VS}	σ_w	β_{BT}	σ_h	W	W, θ

Table 3.2: Summary of the different BLH estimators from peak detection and their notations.

3.1.4 Peak detection recap

As a general conclusion for peak detection methods, a summary of which technique is used on which kind of data is given. Six types of data are available. The table 2.2, page 10, gathers the six estimator, with the method and the data used. From these six data, we create six estimators of the BLH. They will be written $Z(\beta_{VS})$, $Z(\sigma_w)$, $Z(\beta_{BT})$, $Z(\sigma_h)$, $Z(W)$, $Z(W,\theta)$.

- $Z(\beta_{VS})$ is the VS.RciGrad estimator. The BLH is calculated by detecting peak in the Haar wavelet transform of β_{VS} profiles. The default setting is to take the strongest peak, but continuity test can modify it.

- $Z(\sigma_w)$ is the VS.var estimator. The BLH is calculated with the peak-based thresholding method. This method provide only one result, so there is no issue here. Continuity test can nevertheless reject some isolated result.

- $Z(\beta_{BT})$ is the BT.RciGrad estimator. Same as $Z(\beta_{VS})$ but applied on β_{BT} profiles.

- $Z(\sigma_h)$ is the BT.var estimator. Same as $Z(\sigma_w)$ but applied on σ_h profiles.

- $Z(W)$ is the WP.spd estimator. This estimator doesn't use Haar wavelet transform nor peak-based thresholding. The aim of this estimator is to find low-level jet, and the shape of the low-level jet wind profile is not the same as used in the previous techniques. The horizontal wind has already been used to detect boundary layer top by Tucker et al. [2009]. The BLH is thus defined on a peak of the profile, the strongest one is the most likely, but the lowest one could be interesting too in some profiles.

- $Z(W,\theta)$ is the WP.totGrad estimator. Its construction is a bit different from the others in the sense where it uses two types of data. Both W (horizontal wind speed) and θ (horizontal wind direction) are used to calculate u and v, the horizontal component of the wind. Then the Haar wavelet transform is applied on these two horizontal components, and peaks are detected in $\left\|\frac{\partial \mathbf{W}_h}{\partial z}\right\|$.

$$Z(W,\theta) \ : \ \text{first peak in profile of } \left\|\frac{\partial \mathbf{W}_h}{\partial z}\right\| = \sqrt{\left(\frac{\partial u}{\partial z}\right)^2_{\psi_H} + \left(\frac{\partial v}{\partial z}\right)^2_{\psi_H}}$$

where $\mathbf{W}_h = \begin{pmatrix} u \\ v \end{pmatrix}$, is the *total* horizontal wind (hence the name of WP.*tot*Grad), and $\left(\frac{\partial \cdot}{\partial z}\right)_{\psi_H}$ designs the gradient-like Haar wavelet transform.

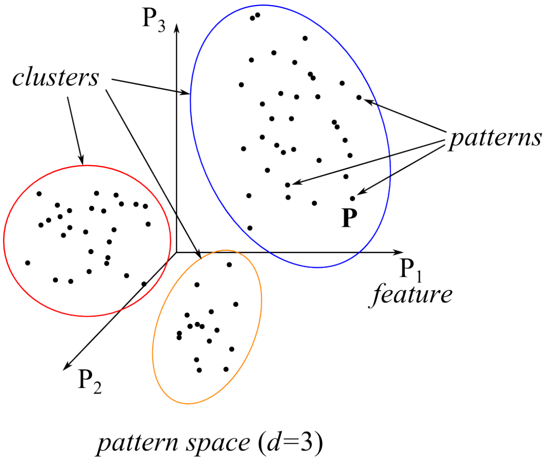


Figure 3.6: Example of *pattern space*, with *patterns*, *features* and *clusters* precised.

3.2 Cluster analysis method

"Cluster analysis" regroups techniques used to classify objects into groups called *clusters*. The purpose of cluster analysis is to create these groups without any prior information about the groups or the hypothetical membership of an object to a group. Two points in the same group will be similar, while two points in different groups will be dissimilar. There are many different ways to do a cluster analysis. A good overview of this domain is provided by Jain et al. [1999].

The first section is a general introduction to cluster analysis where the vocabulary and some useful definitions are stated. The second section focuses on the specific algorithm implemented. The third section discusses about the issue of missing data in such a method.

3.2.1 Concepts and definitions

Cluster analysis is a common tool of data mining. Grouping the data makes it possible to extract useful information from the data set. It is applied in various domains such as pattern recognition (find a specific shape), artificial intelligence (find a relevant among fuzzy input), biology (group similar species), climatology (define a climate), marketing (identify profiles of consumers).

In boundary layer height detection, cluster analysis is a new method. The first application of this method is given by Toledo et al. [2013]. The driven idea is to use cluster analysis to identify the boundary layer on a vertical profile. Using different types of vertical profile (velocity variance, back-scattered intensity...), the common characteristics that correspond to the boundary layer air are tracked in all profiles. Cluster analysis creates groups: one of them is the boundary layer. Looking at the border of these groups, the BLH can be found. This requires homogeneous groups, well separated.

Cluster analysis is now introduced in a more technical way. The following definitions will be used to describe the method with a relevant vocabulary in the next sections.

Definition 3 (*Pattern, pattern space*) *A* pattern **P** *is a single data item, an object that will be assigned into a cluster. It is a vector of d measurements.*

$$\mathbf{P} = \begin{pmatrix} P_1 \\ \vdots \\ P_d \end{pmatrix}$$

A pattern is not modified by the cluster analysis.
All together, the patterns form a pattern set *:* $\mathcal{P} = \{\mathbf{P}^1, \dots, \mathbf{P}^n\}$.
The space of dimension d where belong the patterns is called the pattern space, $\mathbb{P}$.

Definition 4 (*Feature*) *A* feature *is a single scalar component P_j of a pattern* $\mathbf{P}$*. It corresponds to one type of measurement. Each feature is a dimension of the pattern space. Feature space is a synonym of pattern space.*

In order to avoid confusion between feature indexes and pattern indexes, $\mathbf{P}^i$ will indicate the i-th pattern in the pattern set (which is a vector), and P_j will indicate the j-th feature of the pattern $\mathbf{P}$ (which is a scalar). So the j-th feature of the pattern $\mathbf{P}^i$ will be noted P_j^i (scalar).

A pattern $\mathbf{P}$ can also be a point. In the text, I tried to use "point" when it is a measurement (a point in a profile) and "pattern" in the general case. A seed $\mathbf{C}$ is a special pattern that does not come from a measurement. It is the centroid of a cluster (created by the program).

Definition 5 (*Cluster, centroid*) *A* cluster $\mathcal{C}$ *is a group of patterns close to each other in the pattern space and according to the distance chosen. The clusters form a partition of the pattern set : if we have K clusters, then*

$$\bigcup_{k=1}^{K} \mathcal{C}^k = \mathcal{P}$$

The property $\bigcap_{k=1}^{K} \mathcal{C}^k = \emptyset$ is also true, as long as we don't do fuzzy clustering.

For each cluster $\mathcal{C}$, we can define (by different manners) a centroid*. A centroid of the cluster $\mathcal{C}$ is any pattern $\mathbf{C}$ representative of the whole cluster. A synonym for centroid is* seed.

The figure 3.6 shows a *pattern space* with 3 *features*, where several *patterns* gathered into *clusters* are drawn.

For this work, the centroid will be the pattern which each feature is the average feature of the patterns in the cluster.

$$\mathbf{C}^k = \begin{pmatrix} C_1^k \\ \vdots \\ C_d^k \end{pmatrix}, \quad \text{with } C_j^k = \frac{1}{n_k} \sum_{\mathbf{P}^i \in \mathcal{C}^k} P_j^i \ , \ \forall j \in \{1, \dots, d\} \tag{3.4}$$

where n_k is the number of patterns (population) in the cluster $\mathcal{C}^k$ $(1 \leqslant k \leqslant K)$.

On the figure 3.7b, the centroids are the black-circled dots of the color of their cluster. This figure represents a pattern space of dimension 2, from 2 real profiles. In 3.7a are the two profiles (β_{VS} (blue) and σ_w (red), profiles from 01-Apr-2013 22:15:53 UTC) used to create the pattern set visible on 3.7b.

The distance $\delta(\cdot,\cdot)$ used to qualify the proximity of the patterns can be any metric defined on the pattern space. The most common distance is the Euclidian distance:

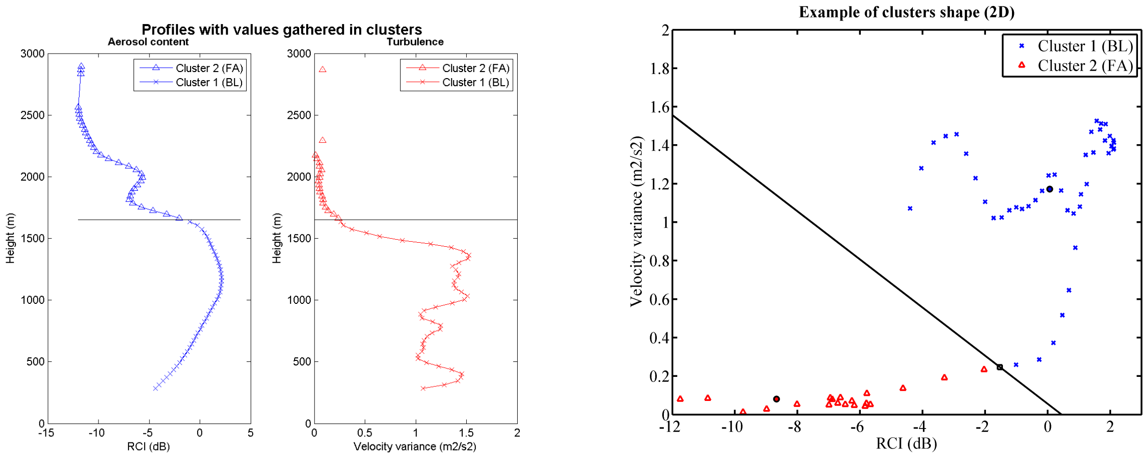
$$\delta(\mathbf{P^j}, \mathbf{P^i}) = \|\mathbf{P^j} - \mathbf{P^i}\| = \sqrt{\sum_{l=1}^{d} (P_l^j - P_l^i)^2}$$

The distance is an important issue in cluster analysis, because it's the metric of dissimilarity that will assess the membership to a cluster. The exact relation between distance and membership may vary in other cluster analysis method. Thus, the next definition is specific to K-means algorithms.

Definition 6 (*Membership*) *Let's consider a pattern set $\mathcal{P} = \{\mathbf{P}^1, \dots, \mathbf{P}^n\}$, and a cluster $\mathcal{C}^k$ among K. The membership $w_k(\mathbf{P}^i)$of the pattern $\mathbf{P}^i$ to the cluster $\mathcal{C}^k$ is*

$$\text{Classic } K\text{-means} \quad w_k(\mathbf{P}^i) = \begin{cases} 1, & \textit{if} \quad \delta(\mathbf{P}^i, \mathbf{C}^k) = \min\limits_{\ell=1\dots K} \{\delta(\mathbf{P}^i, \mathbf{C}^\ell)\} \\ 0, & \textit{else} \end{cases}$$

$$\text{Fuzzy } K\text{-means} \quad w_k(\mathbf{P}^i) = \frac{\delta(\mathbf{P}^i, \mathbf{C}^k)^{\frac{-2}{\phi-1}}}{\sum_{\ell=1}^{K} \delta(\mathbf{P}^i, \mathbf{C}^\ell)^{\frac{-2}{\phi-1}}}$$



(a) Two profiles : β_{VS} (blue) and σ_w (red), interpolated on a common grid to create the pattern set.

(b) Pattern space with the two clusters : boundary layer (blue) and free atmosphere (red). The black line makes the limit visible but is set arbitrarily.

Figure 3.7: Example of real profile clustering

where ϕ is the fuzzifier.
The membership has to verify $\forall i, \quad \sum_{k=1}^{K} w_k(\mathbf{P}^i) = 1$. In classic clustering, this property means that a pattern belongs to only one cluster. In fuzzy clustering, it imposes the membership to be normalized.

The fuzzifier ϕ is a real number that quantifies how fuzzy are the cluster formed. It ranges from 1 to $+\infty$. The closer ϕ is to 1, the sharper are the clusters. The limit case $\phi = 1$ isn't defined but correspond to non-fuzzy clustering ($w_k \in \{0, 1\}$ is logical). Conversely, higher values of ϕ yields to fuzzier clusters and the membership is flat, away from 0 or 1. The limit case $\phi \to +\infty$ leads to undefined clusters with a constant membership $w_k(\mathbf{P}^i) = \frac{1}{K}, \quad \forall\, i,\, k$.

The concept of membership allows a more rigorous expression for the centroid definition than in the equation 3.4. Indeed, the indexation "i / $\mathbf{P}^i \in \mathcal{C}^k$" have to be clarified for fuzzy clustering. Here is the formula to use :

$$\mathbf{C}^k = \begin{pmatrix} C_1^k \\ \vdots \\ C_d^k \end{pmatrix}, \quad \text{with } C_j^k = \frac{\displaystyle\sum_{i=1}^{n} w_k(\mathbf{P}^i) P_j^i}{\displaystyle\sum_{i=1}^{n} w_k(\mathbf{P}^i)}, \ \forall j \in \{1, \dots, d\}$$

This definition is valid for both classic and fuzzy clustering, though classic clustering can be written is a simpler way. Both of them have been used in this work.

Here, the objects that will be manipulated and the relevant vocabulary were presented. The idea of cluster analysis is to form realistic groups and identify one of these groups as the boundary layer. The BLH is then defined as the border of the group. The definition of the membership highlights the difference between classic and fuzzy clustering. Both of them will be used in this work.

3.2.2 Implementation of K-means algorithms

This section will focus on the clustering algorithm chosen: the K-means algorithm, and how it was implemented. The K-means algorithm was chosen for this problem because of the simplicity of its implementation and its robustness ([Jain et al., 1999],[Pollard et al., 1981], [Selim and Ismail, 1984]). The K-means algorithm is a widely used method to gather huge amounts of data into clusters. Since the Doppler lidar can provide high resolution measurements, this method is suitable to process the lidar data. This work presents also an improvement of K-means using fuzzy clustering. The motivation of using fuzzy

K-means is to be able to assess the uncertainty on our BLH estimation. The vocabulary is the same as the previous section.

Let's consider several profiles as a set of n points. For example, profiles from VS data during daytime. A back-scattered intensity profile : ${\beta_{VS}}^1, \dots, {\beta_{VS}}^n$ and a vertical velocity variance profile : ${\sigma_w}^1, \dots, {\sigma_w}^n$, as shown in figure 3.7a. These profiles are interpolated on a common grid to create a pattern set $\mathcal{P} = \{\mathbf{P}^1, \dots, \mathbf{P}^n\}$. The resulting pattern set is on figure 3.7b.

Each pattern $\mathbf{P}^i$, $\forall i \in \{1, \dots, n\}$ is a vector of d variables. In this example, $d = 2$: β_{VS} is one feature, σ_w is the other, and the i-th pattern is the vector $\mathbf{P}^i = \begin{pmatrix} {\beta_{VS}}^i \\ {\sigma_w}^i \end{pmatrix}$. In the K-means algorithm, the number of clusters K is known *a priori*. For each cluster k, $\quad k \in \{1, ..., K\}$, there is one centroid. Clusters are formed around these centroids.

The flowchart of the implemented K-means is given figure 3.8. The input of the program is a raw pattern set $\mathcal{P}_{\text{raw}}$, resulting from the linear interpolation of the data on a common grid (in time and space).

The first box, called "Initializations", define the number clusters, the domain of clustering and the initial membership. In K-means, the number of clusters K is an input parameter. The choice of K could be an issue and it have to be validated afterward. For this work, 2 clusters has been validated as being the best choice (see section 4.2.1).

$$\textit{Number of clusters :} \quad K = 2$$

The domain of clustering is a selection of features and patterns. The selection of features contains the data that was thought to be relevant to use. For example, during the night, because of the minimum range of VS scan, it won't be used. While BT scan is very suitable for shallow boundary layers. So the selection of features will be $\{\beta_{BT}\ ,\ \sigma_h\}$. Conversely, during the day, the maximum range of BT scan is too short, and the BLH is fairly higher than the minimum range of VS. The selection of features will be $\{\beta_{VS}\ ,\ \sigma_w\}$. During the transition period, both are used. Then, the selection of patterns entails ensuring that a data is between the range of the used scan. This selection is transparent for the theory, it's only a computational issue. That is to say that if our scan ranges from 0 to 1000 meters (BT example), only the points of the profile between this gate we be used. The table 3.3 sums up this paragraph.

Time of the day	Night	Morning	Day	Evening
Feature selection	$\{\beta_{BT}\ ,\ \sigma_h\}$	$\{\beta_{VS}\ ,\ \sigma_w\ ,\ \beta_{BT}\ ,\ \sigma_h\}$	$\{\beta_{VS}\ ,\ \sigma_w\}$	$\{\beta_{VS}\ ,\ \sigma_w\ ,\ \beta_{BT}\ ,\ \sigma_h\}$

Table 3.3: Summary of the feature used to cluster depending on the time of the day.

The initial membership is set randomly, along an uniform law. This is a way to avoid sensitivity to initial conditions. After the initialization, the main loop is entered. It is composed with 3 boxes, corresponding to the 3 steps iterative algorithm. A schematic of these steps is given in figure 3.9.

1. **Define the seeds** as the average point in the cluster.
   This step creates a new pattern for each cluster which features are the average feature of all the points in the clusters. If seeds already exists, the old ones are replaced by the new. It corresponds to the "Step 3" in the figure 3.9.

2. **Compute point-to-seed distances**.
   For each point $\mathbf{P}^i$, the distances from $\mathbf{P}^i$ to every seed $\mathbf{C}^1, \dots, \mathbf{C}^K$ are calculated. It corresponds to the "Step 1" in the figure 3.9. It should be noticed that compute none point-to-point distance is computed. This is an important point for section 3.2.3.

3. **Update membership**.
   Here the distances computed at the previous step are used to evaluate the membership of each point. It corresponds to the "Step 2" in the figure 3.9. The formula to calculate the membership from the distances is given on the previous section, with the definition of *membership*.

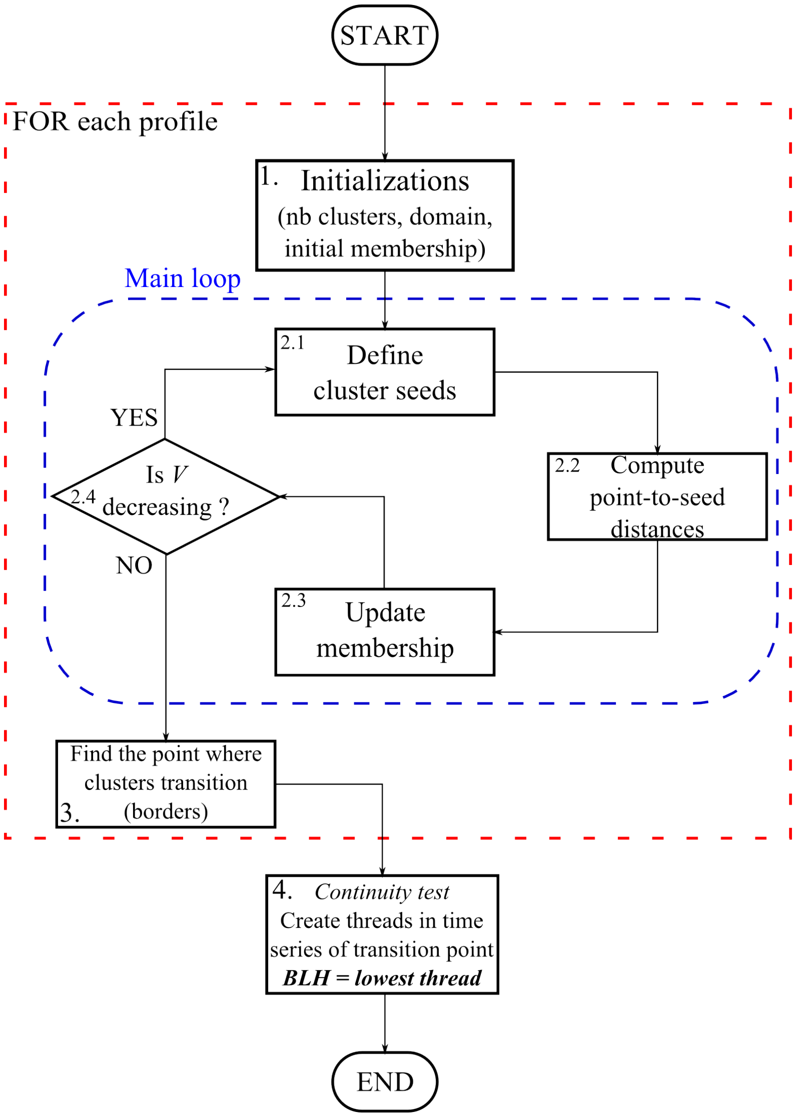


Figure 3.8: Flowchart of the K-means algorithm.

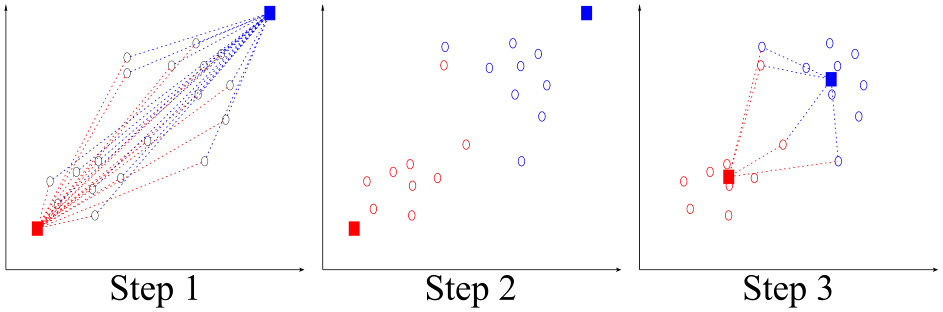


Figure 3.9: Schematic 3 steps in the main loop of the K-means algorithm.
It represents a 2D pattern space, with the patterns (empty circles) and 2 seeds (filled squares). Step 1 shows the "Compute point-to-seed distances". Step 2 is "Update membership", we can see the patterns turning the same color as the closest seed. Step 3 shows the new seeds after the "Define seeds" step, and few point-to-seed distances likely to change the membership at the next step.

4. **Test of convergence**.
   The intra-cluster variance V is tested to check if it is of interest to stay in the main loop. The lower is V, the more the clusters are homogeneous and well separated. The criterion to go out of the loop is the trend of V : while it decreases, the loop continues. But as soon as V increase, the loop is stopped. This ensure a local minimum is reached, which is the best K-means can provide [Selim and Ismail, 1984]. The intra-cluster variance is the sum of squared errors :

$$V = \sum_{k=1}^{K} \sum_{i=1}^{n} w_k(\mathbf{P}^i)\delta(\mathbf{P}^i, \mathbf{C}^k)$$

Then, in the box 3, the clusters are formed, and the job is now to find their border, which is the boundary layer limit. Since the clusters are initialized randomly, it's not possible to say which one is the BL cluster. According to the definition of BL (see 2.2.1, or [Stull, 1988]), the BL cluster is adjacent to the ground. As a consequence, the BL cluster is the cluster that contains the first level of measurement. Instead of testing only the membership of the first level, we test the average membership of the 3 first levels, to avoid possible outlier pattern at the border.

Between the box 3 and 4, the loop of the profile is exited. It's possible to apply the continuity test or not. The reasoning that justifies the continuity test for the peak-based program is still true. In real time, BLH is calculated profile by profile, the program stops here. But when doing reanalysis, continuity test can be applied. The box 4 is the continuity test applied to the points where the cluster transitions.

In summary, the K-means algorithm is an iterative algorithm to create clusters. The cluster are created around centroids called seeds. One of these cluster is the boundary layer, it is assumed to be the one adjacent to the ground. The BLH is then defined as the height where the cluster transitions. The flowchart can be summarized in the following list:

1. Select the useful features. Initialize clusters randomly (uniform law).
2. Enter the main loop (WHILE loop)
   2.1 A seed is defined as the average pattern of a cluster.
   2.2 We compute the distances from each point to each seed.
   2.3 Each point is assigned to the closest cluster (we update the membership).
   2.4 If the intra-cluster variance reaches a minimum, we stops. If it doesn't, go to step 2.1.
3. Find the border of the BL cluster.

It is possible to apply the continuity test to the outputs, as well as for peak detection.

3.2.3 Missing data process

The missing data is a problem in most of areas of research. In this problem, this is critical because it may change the distance we used. For example, in the morning it is very interesting to use information from both BT[4] and VS[5] scans, because the BL is very shallow at the beginning (well seen by BT, out of reach for VS) and grows rapidly until several hundreds meters (well seen by VS, out of reach for BT). But if β_{VS}, σ_w, β_{BT} and σ_h are used as features, our pattern space is 4 dimensional, and the distance is defined by $\mathbb{R}^4$. The high points in the profile where VS is available but BT is not are patterns where 2 features are missing. The correct pattern space of these pattern is $\mathbb{R}^2$. Which distance should be used to evaluate the dissimilarity of a pattern out of BT/VS reach? Can we compare it to the other patterns?

These questions are essential to make sure that the cluster analysis method is consistent. The figure 3.10 represents an example of the problem on a theoretical 3D pattern space. The example is given by

[4] (Reminder) BT : "BowTie" partial RHI scan, see figure 2.2, page 7.

[5] (Reminder) VS : "Vertical Staring", see figure 2.3, page 7.

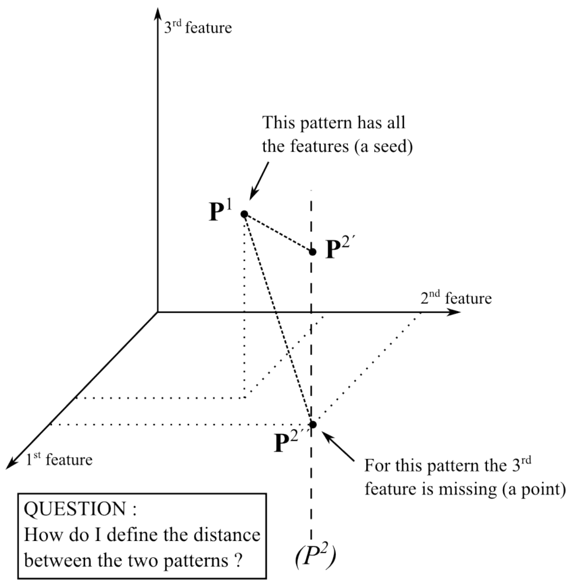


Figure 3.10: Missing data problem : a 3D feature space with 2 patterns are represented. The pattern $\mathbf{P}^1$ has all the features available, we are able to position it correctly in the feature space. The pattern $\mathbf{P}^2$ has the 3rd feature missing, it's only possible to affirm he belongs to the vertical dashed line. Two possible choices of position for $\mathbf{P}^2$ are shown : 3rd feature set to zero ($\mathbf{P}^{2''}$) and 3rd feature set to the same as $\mathbf{P}^1$ ($\mathbf{P}^{2'}$).

two patterns : pattern $\mathbf{P}^1$, with all the features known, and pattern $\mathbf{P}^2$ with the last feature unknown. For the pattern $\mathbf{P}^1$, there is only one location possible in the pattern space. But for the pattern $\mathbf{P}^2$, we cannot locate it on the 3rd axis, because the 3rd feature is unknown. It's only possible to say it belongs to the vertical dashed line, but not where. Trying to assess the dissimilarity between these two points, the result will depend on the choice for the processing of the missing data.

Definition 7 (*Incomplete pattern*) *A pattern with missing feature is called* incomplete.
For any pattern $\mathbf{P} = \begin{pmatrix} \mathbf{P}_1 \\ \vdots \\ \mathbf{P}_d \end{pmatrix}$*, we define the set of features available :*

$$R_{\mathbf{P}} = \{j \ / \ P_j \text{ exists}\}$$

The distance used on incomplete pattern is not properly a distance, so we will call it pseudo-distance *or* incomplete distance*,* $\tilde{\delta}$*.*

The figure 3.10 presents two possible choices of incomplete distance.

First choice is to set the missing value to a given value (0 for example). The resulting pattern $\mathbf{P}^2$ is located on $\mathbf{P}^{2''}$ in this case. This solution is simple and has the advantage to take into account, in a questionable extent, the variability along the missing feature of the full known pattern. But the disadvantage is obviously the choice of the given value. There is no reason for 0 (or whatever given value) to be better than another one. The pseudo-distance used here is

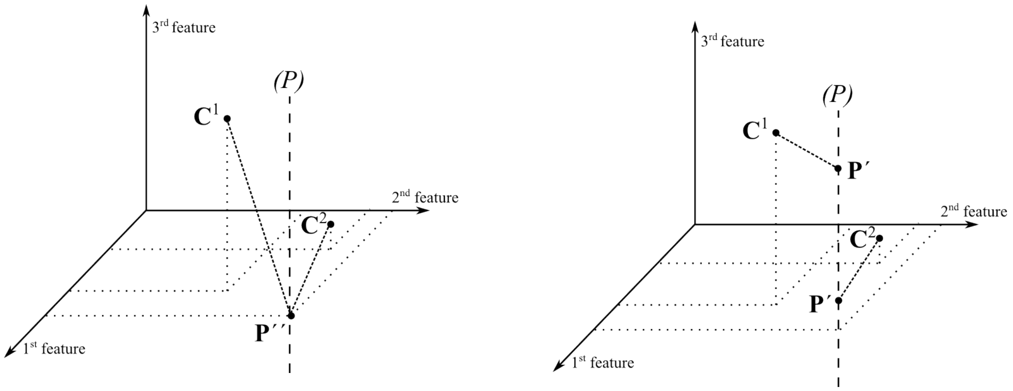
$$\tilde{\delta}_\alpha(\mathbf{P}^1, \mathbf{P}^2) = \delta(\mathbf{P}^1, \mathbf{P}^{2''}) = \sqrt{\sum_{l \in R_{1,2}} (P_l^1 - P_l^2)^2 + \sum_{l \notin R_{\mathbf{P}^1}} (\alpha - P_l^2)^2 + \sum_{l \notin R_{\mathbf{P}^2}} (P_l^1 - \alpha)^2}$$

where $R_{1,2} = R_{\mathbf{P}^1} \cap R_{\mathbf{P}^2}$ and α is the imputation value.

Second is define the distance between $\mathbf{P}^1$ and $\mathbf{P}^2$ as the distance between the pattern $\mathbf{P}^1$ and the dashed line. This is equivalent to taking the same 3rd feature value for $\mathbf{P}^1$ and $\mathbf{P}^2$. The resulting pattern $\mathbf{P}^2$ is then located on $\mathbf{P}^{2'}$. The advantage is to be less arbitrary than set a given value instead of the missing data. But the shortcoming is to neglect the variability along the missing feature of the full known pattern. Another drawback is to give artificially short distances when a data is missing (because of the number of terms in the sum). The pseudo-distance used here is

$$\tilde{\delta}_A(\mathbf{P}^1,\mathbf{P}^2)=\delta(\mathbf{P}^1,\mathbf{P}^{2\prime})=\sqrt{\sum_{l\in R_{1,2}}(P_l^1-P_l^2)^2}$$

where $R_{1,2}=R_{\mathbf{P}^1}\cap R_{\mathbf{P}^2}$.



(a) Imputation with given value (here 0) : pseudo-distance $\tilde{\delta}_0$

(b) No imputation : pseudo-distance $\tilde{\delta}_A$

Figure 3.11: Different incomplete distances used with missing data.
Here is a theoretical example with one pattern $\mathbf{P}$ compared to two seeds $\mathbf{C}^1$ and $\mathbf{C}^2$ in a theoretical 3 dimensional pattern space.

The important point in this discussion is to understand how the way the missing data are processed influences the final result. The K-means algorithm compares each point of the profile to each seed. We make sure that each seed has no missing feature. But the profiles are given by the lidar, and missing data are likely. Considering one point, it is assigned to a cluster by comparing the distances to all the seeds (see membership definition). The figure 3.11 shows an example of one point, compared to two seeds. The two pseudo-distances described before are shown ($\tilde{\delta}_0$ is 3.11a, $\tilde{\delta}_A$ is 3.11b). If a feature is missing for a point $\mathbf{P}$, it will affect the distance from $\mathbf{P}$ to all the seeds *in the same way*. For example, considering that $\mathbf{P}$ is a pattern with one missing feature. We will assign $\mathbf{P}$ to a cluster by comparing the distance to every seed $\mathbf{C}^1,\ldots,\mathbf{C}^K$. It's impossible to compute the complete distance $\delta(\mathbf{C}^k,\mathbf{P}),\ k\in\{1,\ldots,K\}$ because of the missing feature in $\mathbf{P}$. Instead we use one of the pseudo-distance $\tilde{\delta}(\mathbf{C}^k,\mathbf{P})$ described before. So we will compare $\tilde{\delta}(\mathbf{C}^1,\mathbf{P})$ to $\tilde{\delta}(\mathbf{C}^2,\mathbf{P})$: the pseudo-distance $\tilde{\delta}$ is the same.

Point-to-seed incomplete distances can be compared as long as the point is the same for all the pseudo-distances compared. But we cannot compare point-to-seed incomplete distance with different points, even if the seed is the same. Indeed, the point-to-seed pseudo-distance $\tilde{\delta}(\mathbf{C},\mathbf{P})$ depends on the point $\mathbf{P}$. The number of feature could be different from one point to another, and it modifies the pseudo-distance. That is to say, for a given pattern $\mathbf{P}$ and two seeds $\mathbf{C}^1$, $\mathbf{C}^2$, it is correct to write

$$\tilde{\delta}(\mathbf{C}^1,\mathbf{P})\leqslant\tilde{\delta}(\mathbf{C}^2,\mathbf{P})\quad\text{for example}$$

But it is not correct if the point is not the same. For the same seed $\mathbf{C}$ and two points $\mathbf{P}^1$, $\mathbf{P}^2$, the distance is not comparable.

$$\tilde{\delta}(\mathbf{C},\mathbf{P}^1)\not\leqslant\tilde{\delta}(\mathbf{C},\mathbf{P}^2)\quad\text{for example}$$

For $\tilde{\delta}_\alpha$, the effect will depend on the α value. For $\tilde{\delta}_A$, the distance value will increase with the number of feature available.

A more complete process of the missing data is to make data imputation. Imputation is the process of replacing a missing data by value. Methods exist to choose a relevant value according to the data available. Li et al. [2004] provided a method using fuzzy clustering to make data imputation. The implementation of cluster analysis used in this work could be used to make data imputation.

4 | Results and discussion

4.1 A case study

The strengths and weaknesses of each method will be investigated for a particular day: the 1$^{\text{st}}$ of April 2013. This day is characterized by clear sky, though some convective activity is visible in the afternoon. I chose this day for the case study because there are very few missing data. The data as measured by the lidar are shown figure 4.1.

Figure 4.1 shows the six data measured by the lidar (described on the section 2.1). The table 2.2, page 10, provides concise information about it. The figures 2.1 to 2.3, page 7, complete it.

The convective boundary layer is easy to see during the daytime (the red line is the solar radiation), especially in the data from vertical staring, β_{VS} (VS.RCI) and σ_w (VS.var). The increasing boundary layer depth is visible in the data from BT scan too, but the maximum height of BT is reached quickly. In the first two panels (horizontal wind speed and direction), the shape the boundary layer height is invisible. Only the shear in wind direction during the day seems to correctly track the boundary layer height. As a consequence, we might not expect accurate results from the wind information for this day. An interesting thing to notice is the vertical layered structure of the wind during the night, that supports the discussion in section 2.2.2.

During the night, the transition is less obvious. It is observed in σ_h (BT.var) as a very shallow layer with a little more turbulence. It is present in β_{BT} (BT.RCI) too, with even less contrast though. The minimum range of VS doesn't allow to find any shallow transition. A better way to see it is to look directly at the profiles. The figure 2.7a (page 12) represents the profiles from the bowtie scan at 06:29 UTC. The shallow layer is more visible here, although we can still discuss whether it is correct to call that layer a "boundary layer". Since our purpose is essentially air quality and horizontal fluxes, we will consider it as the boundary layer, because that the layer where the emitted (from the ground mostly) aerosols will remain. Our methods have been developed in order to catch such layers.

From this visual examination of the raw data, the boundary layer height seems to be very shallow during the night (around 200 m), start rising around 16:00 UTC, and reaches rapidly the top of the convective boundary layer (around 1500 m). The methods are expected to match with this scheme to be considered as valid.

4.1.1 Peak detection

The first result shows the described methods applied directly on the data, with thresholding the peaks nor applying the continuity test. The settings for these results are given table 4.1.

To present these first results, the same figure as 4.1 will be used. The BLH estimation derived from a particular type of data will be shown as black dots over the original data. This choice of display allows us to check if the estimations match with the original data. This image is the figure 4.2. The second graph exploited in the next section are peak maps. On the x-axis, one time index corresponds to one profile. The time resolution of the lidar is 20 minutes. On y-axis is the height. On each profile there are several peaks, at different heights, shown by grey dots.

No thresholding, no continuity test (figures p. 33 to 34). The figure 4.2 will be described from the top to the bottom. Because it is the first time, we will provide more details on the description.

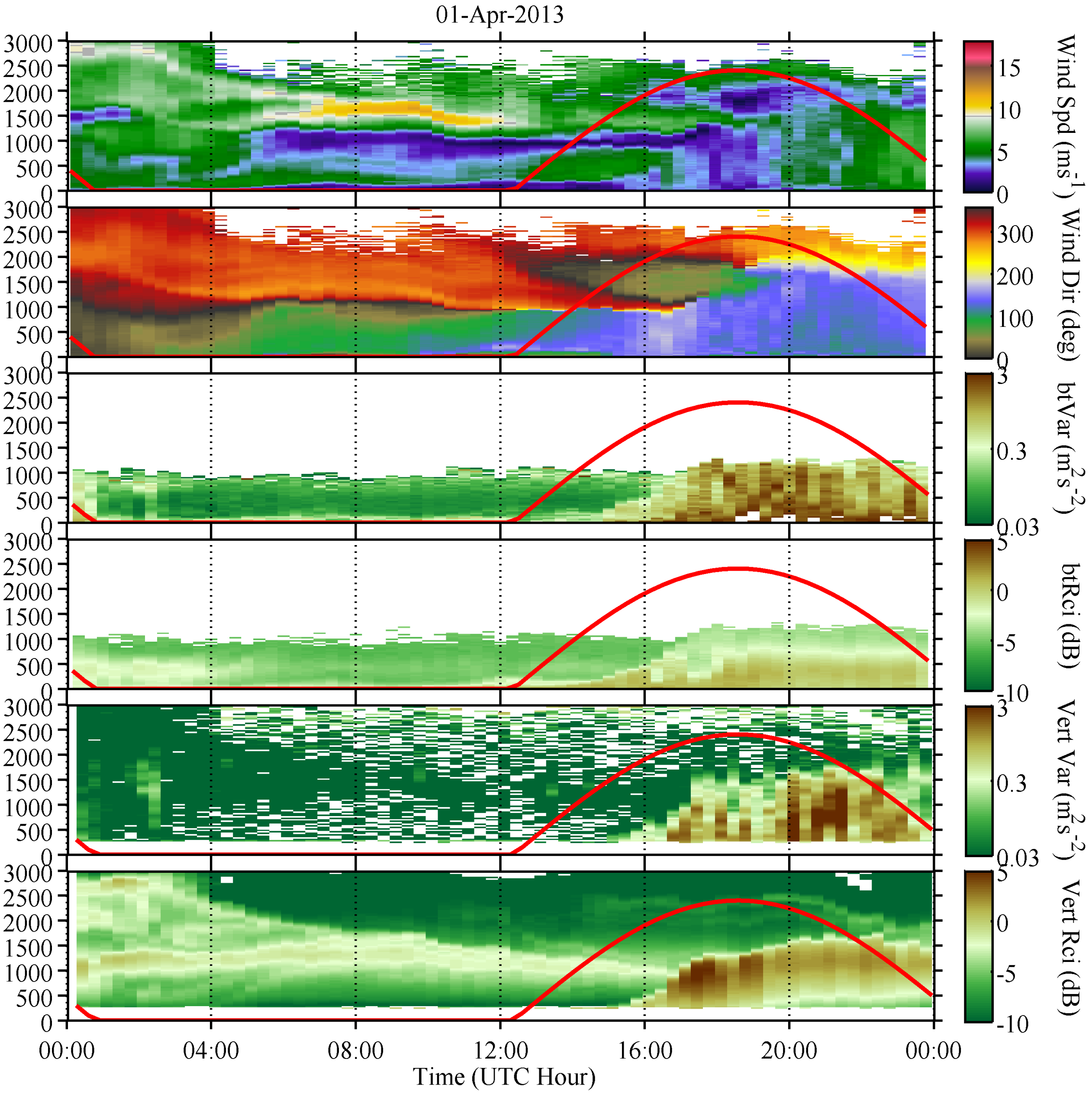


Figure 4.1: Original data for the 01-Apr-2013.
Each panel corresponds to a single type of data in shades of colors. From the top to the bottom : W (horizontal wind speed), θ (horizontal wind direction), σ_h (horizontal velocity variance), β_{BT} (RCI), σ_w (vertical velocity variance), β_{VS} (RCI). In x-axis is the time of the day (UTC), y-axis is the height (from 0 to 3000 m). The red solid line is the theoretical solar radiation.

- ⋆ The first panel is the $Z(W)$ BLH estimation, with W field in shades of colors in background. The first observation is that the boundary layer shape is invisible in the wind speed field. As a consequence, the estimation is not good. The interesting thing to notice, that the dots follow the little jet at 1500 m between 6:00 UTC and 10:00 UTC. It's not a low level jet properly because it's too high, but it's good that the program detects it. On the corresponding peak map (fig. 4.3a), we can see that there are many other peaks that could poison the result, but the dots still follow what looks like a LLJ. Since there is no clear structure in wind speed that could shape the boundary layer (like low-level jet, or wind shear at the inversion level), there is not many to expect from this estimator.

- ⋆ The second panel is the total wind gradient estimator $Z(W, \theta)$, with the wind direction θ in shades of colors in background. We can make the same comment as for the previous panel : there is no wind structure that could identify the boundary layer, so this estimator will not be very good. Although, we can see in the wind direction field a shear at the top of the BL during the day. But the dots of $Z(W, \theta)$ don't follow it. To check if the problem comes from the detection or the peak selection, the corresponding peak map, figure 4.3b, can help. There are so many peaks that it's impossible to perceive any structure. By thresholding the peaks, we would expect to keep only the significant peaks, but with the current settings, this estimator doesn't work.

- ⋆ The third panel is the horizontal velocity variance estimator $Z(\sigma_h)$, with σ_h in shades of colors in background. Because the data come from bowtie scan, which is known to have a very low maximum height (see summary page 7), the data are sparse above 1000 meters. Although, during the night, $Z(\sigma_h)$ tracks the shallow BL very well. On the corresponding peak map, figure 4.3c, we can see that the estimation doesn't follow the peaks. This is because for velocity variance, we use the peak-based thresholding method. And in this method the peaks are used to set the threshold value. The transition point is the highest point in the profile above the threshold, regardless if it's a peak or not. That's why the blue line doesn't follow the grey spots. Another thing: the estimation stops around 16:00 (time index no. 50), even though data is available (according to fig. 4.2) and peaks are detected (according to fig. 4.3c). This occur because no point in the bowtie range meet the threshold, or because there is no peak connected to the ground. In this case, the first explanation is more likely.

- ⋆ The fourth panel is the range-corrected intensity from BT scans β_{BT} in background, with the associated estimator $Z(\beta_{BT})$. It is the same scan as the previous panel, so it has the same bounds on data availability. Also it has the same skill to observe the shallow BL, though we can notice several differences. The first difference is that the estimation doesn't stop after the morning transition. Here we use the Haar wavelet transform method, and it will provide something as long as there is a peak in the gradient of β_{BT}. As we can check on the corresponding peak map (fig. 4.3d), there are peaks in every single profile of the day. Since we choose the strongest peak, regardless of its absolute value, it is likely that the chosen peak after the morning transition (rise before 16:30 UTC/time index 50) is not the BL top anymore. The thing to remember is don't trust $Z(\beta_{BT})$ during the day.

- ⋆ The fifth panel is the vertical velocity variance σ_w in background, with its associated estimator $Z(\sigma_w)$. For vertical staring, the range limitation is the opposite of bowtie: it is the minimum range of about 300 meters that is the limiting factor. As a consequence, there are almost no dots during the nighttime. Only a few dots exist - most likely these are outliers. During the day conversely, the turbulence, visible in color shades, is well tracked. The height of the dots is quite variable though, which could make it difficult to apply the continuity test. We can make the same remark as for $Z(\sigma_h)$ (because it is also the peak-based thresholding applied), the blue line on the peak map figure 4.3e doesn't follow the peaks for the same reason. The thing to remember is that $Z(\sigma_h)$ is good during the day, and most of the time, it provides nothing when it isn't expected to work (nighttime).

- ⋆ The sixth and last panel is the estimator $Z(\beta_{VS})$, with the original data β_{VS} in background. For this data, we have noticed that the transition zone is very visible, especially during the day. We can see that the dots follow this transition pretty well. But during the nighttime, an estimation is still provided, for every single profile. This nighttime estimation jumps between the minimum range of the data and what looks like a residual layer. None of them are good, because the first one seems to be due to the end of the data and not to any shallow BL, and the second one is not the phenomenon we want to catch. The peak map (fig. 4.3f) shows the peak the β_{VS} gradient. This map, like the $Z(\beta_{BT})$ one, is clearer than the velocity variance ones. The thing to remember is that $Z(\beta_{VS})$ is a very good estimator during the day, but shouldn't be trusted during the night.

To sum up this paragraph, $Z(W)$ is not of interest when there is no LLJ nor wind gradient at the BL top. $Z(W,\theta)$ could be able to catch the wind shear that exists during daytime, but suffers from too many detected peaks. $Z(\sigma_h)$ and $Z(\beta_{BT})$ are both good at night, but they are not able to reach the BLH during the daytime. $Z(\sigma_h)$ has the advantage to stop when it is wrong, while $Z(\beta_{BT})$ keeps giving something which is not BL top. Estimators from vertical staring have the same properties as bowtie scan ones. The difference is the period when they work, due to the opposite limitation of range. They are good during daytime and bad at night, and $Z(\beta_{BT})$ keeps giving wrong results while $Z(\sigma_w)$ stops.

Peak thresholding, no continuity test (figures p. 36). For this discussion, the emphasis will be put on the cases that we identified to be problematic without thresholding : $Z(W,\theta)$, $Z(\sigma_h)$ and $Z(\sigma_w)$. The threshold values have been set after a similar discussion, but it is not presented here. They are summarized in the table 4.2, where the modified rows are flagged in cyan.

Looking at the figure 4.4, which is the same as 4.2 with only the panels (θ , $Z(W,\theta)$), (σ_h , $Z(\sigma_h)$) and (σ_w , $Z(\sigma_w)$), not much improvement is seen in the total wind gradient estimator $Z(W,\theta)$. The two others are unchanged. To explain these assessments, the peak maps can help (figure 4.5). Here differences are visible.

For the total wind gradient, figure 4.5a, the improvement is clear, compared to the peak map without peak thresholding (fig. 4.3b). Instead of the freckled map without any structure, we have a clearer map where we can distinguish the shape of the BLH. Even during the night, we can see several spots that look like the top of the shallow BL. But the blue line, which corresponds to the black dots on the data, doesn't follow the good peaks. The detection method seems to be good, with the filtering of small peaks, but the selected peak is not the good one.

For the two others, $Z(\sigma_h)$ and $Z(\sigma_w)$, we note the same effect of cleaning up the map. But if it is like the BL is cleaned up in the $Z(W,\theta)$ peak map, on the figure 4.5b and 4.5c, it's more like the free atmosphere was cleaned up. Indeed, we have seen in the figures 2.7a and 2.7b that the highest peak on velocity variance are in the BL (because of turbulence). So, the clean up is not surprising, and the shape of the freckles show that the value is well chosen. As we have seen in figure 4.4, the final result is unchanged. So, the thresholding is not an issue for these estimators, but the threshold value is. Indeed, these estimators come from the peak-based thresholding method, and though the method is built to avoid the threshold-dependency, it still depends on the threshold value. This peak map with thresholding is a way to check that the threshold is well chosen.

To conclude with the peak thresholding effect, we can say that it cleans up the useless peaks that may poison the results, but it doesn't improve the results alone. The way we select the peak have to be investigated. It will be discussed through the continuity test assessment. Indeed, the choice of the thread is an important parameter in the continuity test.

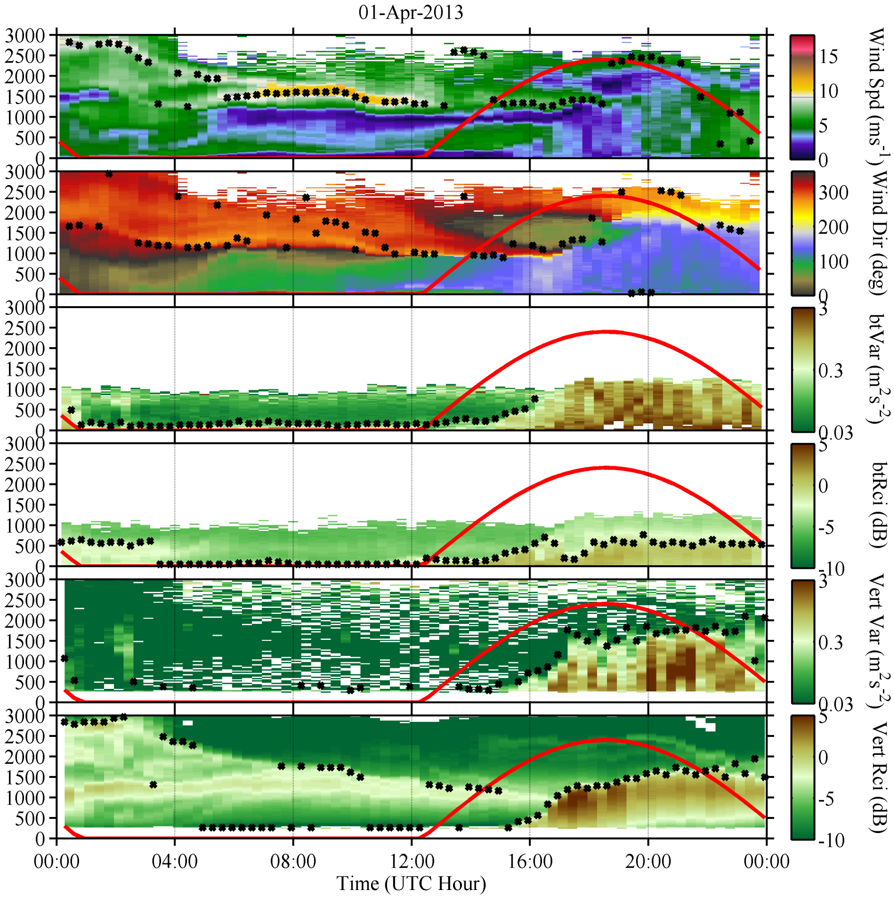
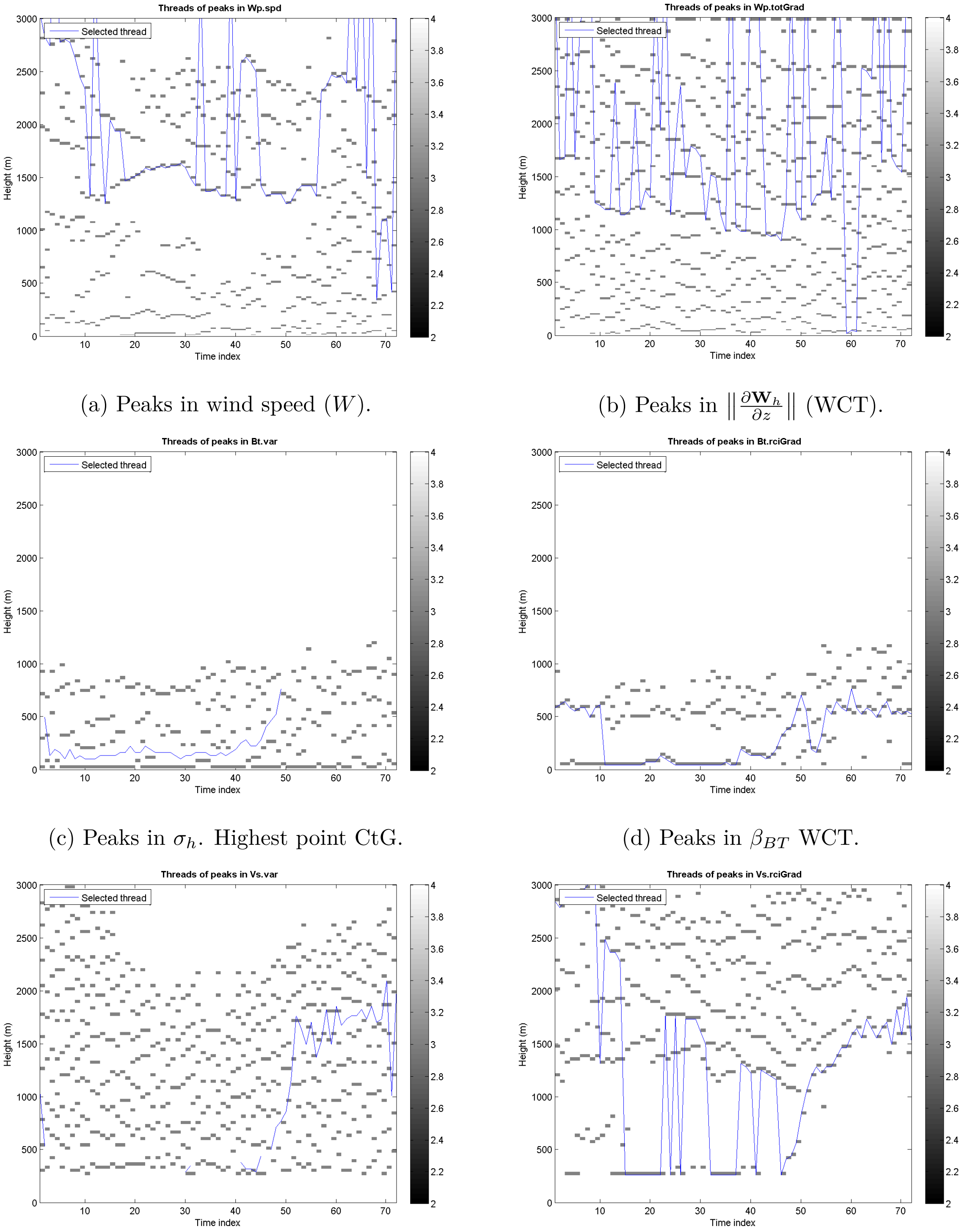


Figure 4.2: Results without thresholding and without continuity test. The table of settings is 4.1.

Estimator / Parameter	$Z(\beta_{VS})$	$Z(\sigma_w)$	$Z(\beta_{BT})$	$Z(\sigma_h)$	$Z(W)$	$Z(W,\theta)$
Type of estimation	Haar wavelet transform	Peak-based thresholding	Haar wavelet transform	Peak-based thresholding	Peak in the profile	Haar wavelet transform
Threshold peaks?	no					
Apply continuity test?	no					
Chosen peak	Strongest					

Table 4.1: Table of settings no. 1. Recap settings of the peak detection program for each estimator. The estimators are named according to table 3.2, page 20.



(a) Peaks in wind speed (W).

(b) Peaks in $\left\| \frac{\partial \mathbf{W}_h}{\partial z} \right\|$ (WCT).

(c) Peaks in σ_h. Highest point CtG.

(d) Peaks in β_{BT} WCT.

(e) Peaks in σ_w. Highest point CtG.

(f) Peaks in β_{VS} WCT

Figure 4.3: Peaks maps without threshold peak, nor continuity test. Selected peak is the strongest. The color bar is arbitrary: grey spot = 1 peak, white = nothing.

4.1.2 Continuity test

Now the continuity test will be discussed. It is described in section 3.1.3. The idea is that the boundary layer top is mostly continuous, and varies quite slowly. The aim of this test is to track the spatial and temporal continuity of the peaks visible on the peaks maps (fig. 4.3 and 4.5). The test depends on three parameters :

- Δz : the width in height of the test window.
- Δt : the width in time of the test window.
- The chosen thread (could be *longest*, *lowest*, *strongest*).

We will start the by discussing the default parameter values, examine revelant modifications, and present the final settings.

Default continuity test : $\Delta z = 3$, $\Delta t = 2$, *longest* thread (p. 39 to 40). This discussion is presented like the two previous : on page 39, there are the estimations over the data in color shades (fig. 4.6) and the table of settings (table 4.3). On page 40, there are the peaks for every estimator. Because we are applying continuity test, the peaks are now gathered in threads, and the color represents their length. The figure 4.6 will be described from the top to the bottom.

⋆ $Z(W)$: Compared to the first result (fig. 4.2), we can see the effect of the continuity test. Instead of tracking the maximum wind, the estimator follows the longest thread (as we set it). The good thing is that this layer seems closer to BL top, but we have to question the reality of this thread. As we have seen before (fig. 4.2), is it quite easy to detect a peak in a profile. Even though the threshold gets rid of the less significant ones, it is likely that some peaks that are not a LLJ are above the threshold. The stable conditions at night might let these peaks stay long enough to make a long thread. In conclusion, the choice of the longest thread for $Z(W)$ is not relevant. A better choice is the strongest peak.

⋆ $Z(W, \theta)$: This estimator isn't improved either by this use of the continuity test. As we can see on figure 4.6, the estimator doesn't follow the shear at the top of the convective boundary layer. On the figure 4.7b, the spots that tracked the shear feature have disappeared. They are probably not smooth enough to get caught by the continuity test. Here, this is the width of the window that should be modified. A larger window should be better.

⋆ $Z(\sigma_h)$: For σ_h, we use the peak-based estimator. That means that we don't choose a peak, but the highest point connected to the ground. The continuity test for these estimator is not applied on the peaks but on the estimate itself. For a given estimate, we look into the window, and if we find another estimate, they are linked in a thread, just like it's done for peaks. The difference is that there is only one peak-based estimate by profile while there are several peaks. So the choice is limited to keep the estimation (if it's in a thread) or leave it (if it isn't in a thread). We can see that $Z(\sigma_h)$ is quite continuous and so, not too filtered by continuity test.

⋆ $Z(\beta_{BT})$: Here too, the continuity test doesn't improve the results. The estimator doesn't catch the shallow BL anymore. A look at the figure 4.7d indicates that the longest thread is the layer above, though it is less smooth than the shallow layer. In this case, the window is too wide because the continuity test shouldn't keep this irregular thread so long.

⋆ $Z(\sigma_w)$: For the same reason as the σ_h, the continuity test is a bit different. It's applied on the estimator itself and not on the peaks. And described previously (fig. 4.2), this estimator can have extended variation in the vertical. As a consequence, the continuity test as it is set doesn't catch it and filters it out. That's why we have only few dots on the picture. A wider window along height might improve the results.

Parameter \ Estimator	$Z(\beta_{VS})$	$Z(\sigma_w)$	$Z(\beta_{BT})$	$Z(\sigma_h)$	$Z(W)$	$Z(W,\theta)$
Type of estimation	Haar wavelet transform	Peak-based thresholding	Haar wavelet transform	Peak-based thresholding	Peak in the profile	Haar wavelet transform
Threshold peaks?	yes					
Threshold value	0.1	0.5	0.1	0.2	3	0.75
Apply continuity test?	no					
Chosen peak	Strongest					

Table 4.2: Table of settings no. 2 (thresholding on, continuity test off). The estimators are named according to table 3.2, page 20.

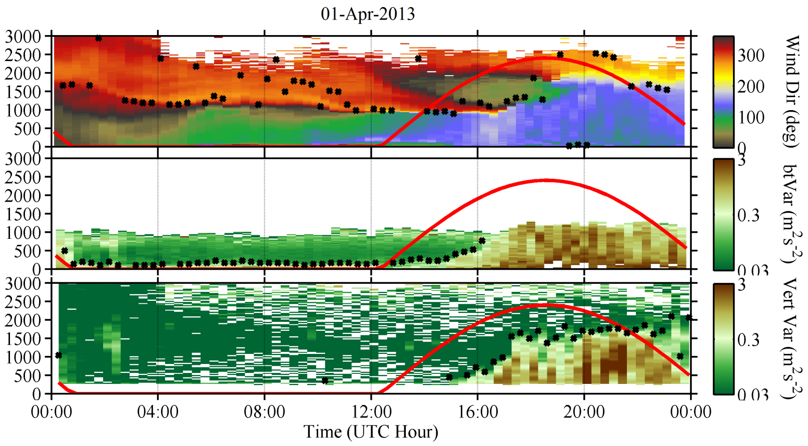
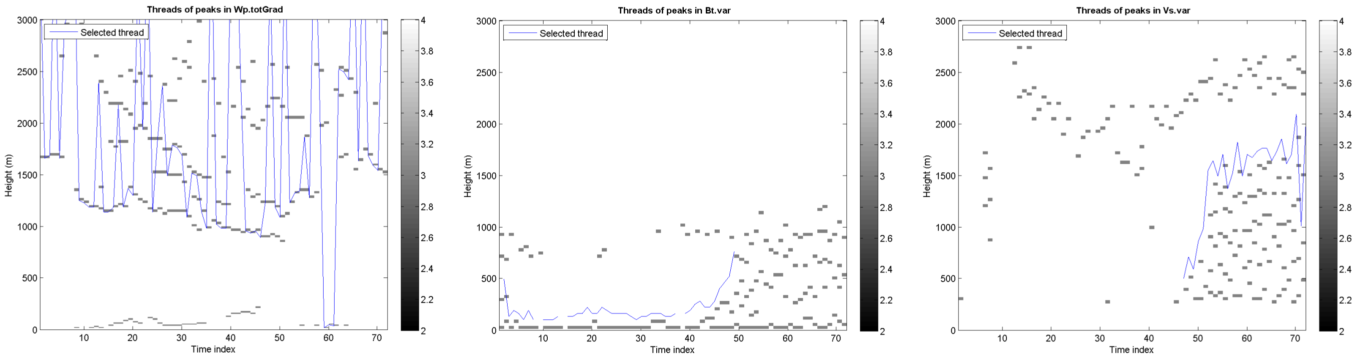


Figure 4.4: Results *with thresholding* and without continuity test. The table of settings is 4.2.



(a) Peaks in $\left\|\frac{\partial \mathbf{W}_h}{\partial z}\right\|$ (WCT). (b) Peaks in σ_h. Highest point CtG. (c) Peaks in σ_w. Highest point CtG.

Figure 4.5: Peaks maps *with peak thresholding*, but without continuity test. Selected peak is the strongest. The color bar is arbitrary: grey spot = 1 peak, white = nothing.

- ⋆ $Z(\beta_{VS})$: The comparison to the data doesn't look good. When we look at 4.7f, we can identify several things: the top of the convective boundary layer during daytime, one or two residual layers during nighttime, and the peak stuck close to the minimum range. The morning rise of the BL has been filtered by the continuity test (we can see it on fig. 4.3f). During the night, we have concluded that this estimator shouldn't be trusted. Since we won't use it during the night, we can afford to keep the lowest thread during the night (even it is an artificial one) in order to include the morning transition in the thread.

As a conclusion for this paragraph, the current settings of the continuity test don't improve the results. The reasons are mostly the settings of the window, which should be customized for each estimator. In some cases ($Z(\beta_{VS})$, $Z(W)$) the problem is the choice of the thread. We will now try the modifications suggested by the observation.

Customized continuity test (p. 41). This last try uses customized values shown in table 4.4, and suggested by the observation in the last paragraph. The results are in figure 4.8, page 41. The values have been adjusted by repeating the same process of visual examination and critique of the results.

The improvement is mostly visible on the total wind gradient estimator $Z(W, \theta)$. The first panel of 4.8 shows $Z(W, \theta)$ following the wind shear better than before the continuity test. There are still some profiles during the day where the dots aren't on the shear feature, but for these dots, this more a drawback of the method than a drawback from continuity test.

For the others results, the improvement of continuity test is more tempered. Indeed, if we compare with the first results on figure 4.2, (p, 33), the difference isn't obvious. This is because the 1^{st} of April 2013 is a day without too much problems: few missing data, clear sky, fair diurnal cycle. The continuity test is a good filter for outliers. On this day, there is not that much to filter, that is why the difference is so minimal. But on other days, it might be different. And nevertheless, we can note that the isolated dots on the $Z(\sigma_w)$ estimation disappeared, which is an example of the filtering effect of the continuity test.

In summary, the continuity test is a post-process of the estimation that looks for continuity among the estimates. Sometimes it's a very helpful process because it helps finding the good transition. An example was provided here with the total wind gradient estimator $Z(W, \theta)$. Most of the time, it is only a filter to remove isolated value likely to be outliers. This case study is obviously not enough to really assess the benefit of the continuity test. The reasoning presented here have to repeated on different days, with different characteristics.

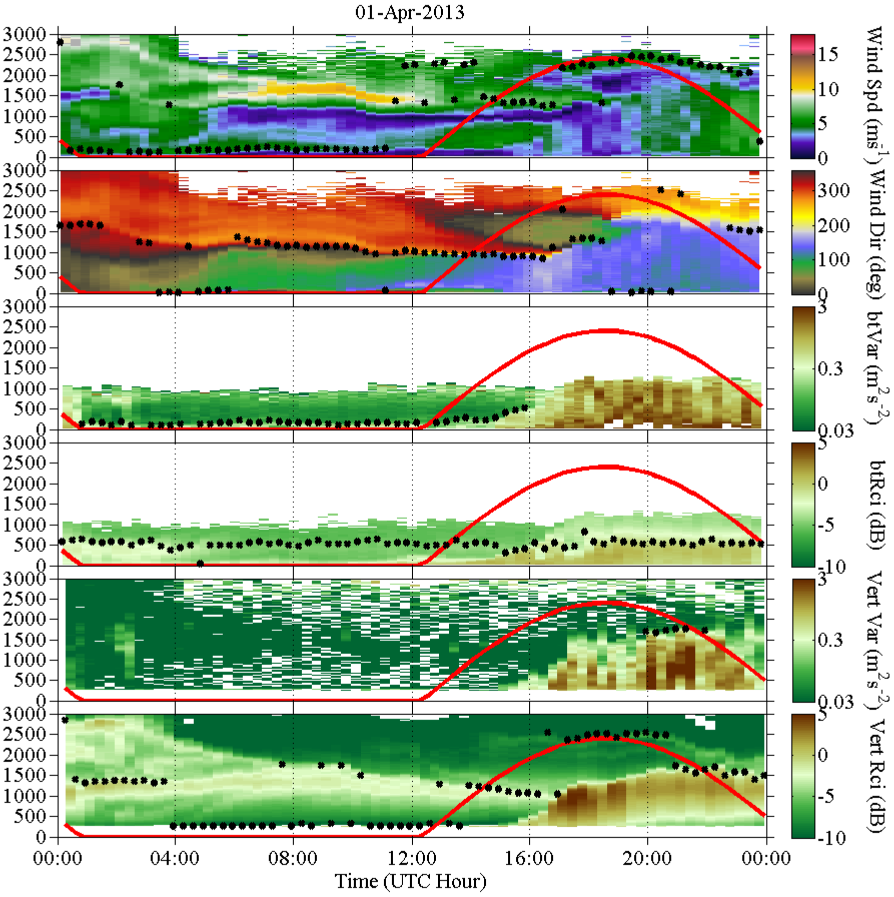
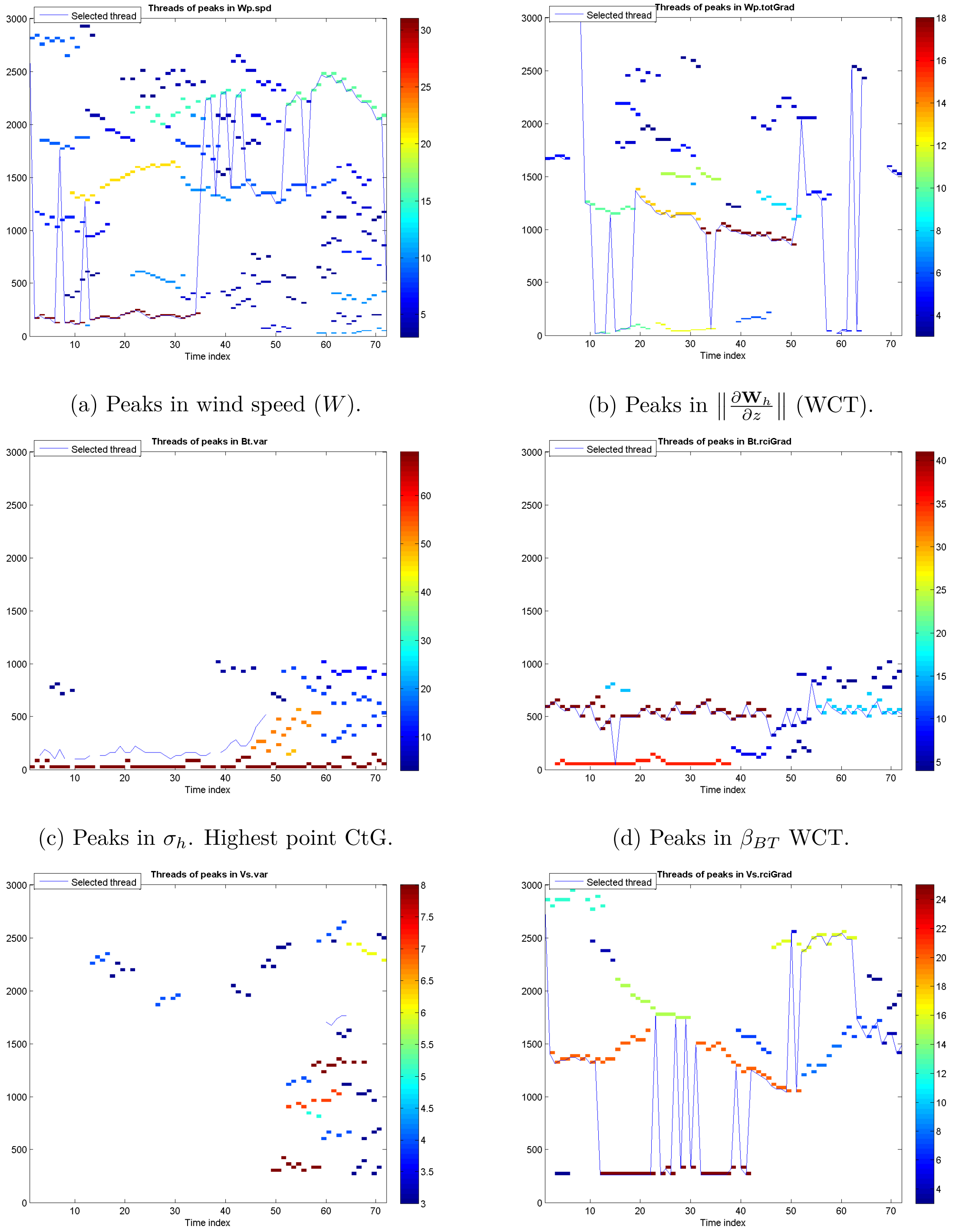


Figure 4.6: Results without thresholding and without continuity test. The table of settings is 4.3.

Estimator / Parameter	$Z(\beta_{VS})$	$Z(\sigma_w)$	$Z(\beta_{BT})$	$Z(\sigma_h)$	$Z(W)$	$Z(W,\theta)$
Type of estimation	Haar wavelet transform	Peak-based thresholding	Haar wavelet transform	Peak-based thresholding	Peak in the profile	Haar wavelet transform
Threshold peaks?	yes					
Threshold value	0.1	0.5	0.1	0.2	3	0.75
Apply continuity test?	yes					
Δz	3					
Δt	2					
Chosen peak	In longest thread					

Table 4.3: Table of settings no. 3. Recap settings of the peak detection program for each estimator. The estimators are named according to table 3.2, page 20.



(a) Peaks in wind speed (W).

(b) Peaks in $\left\|\frac{\partial \mathbf{W}_h}{\partial z}\right\|$ (WCT).

(c) Peaks in σ_h. Highest point CtG.

(d) Peaks in β_{BT} WCT.

(e) Peaks in σ_w. Highest point CtG.

(f) Peaks in β_{VS} WCT

Figure 4.7: Peaks maps with default continuity test. Selected peak is in the longest thread. The color bar is the thread length (in number of spots).

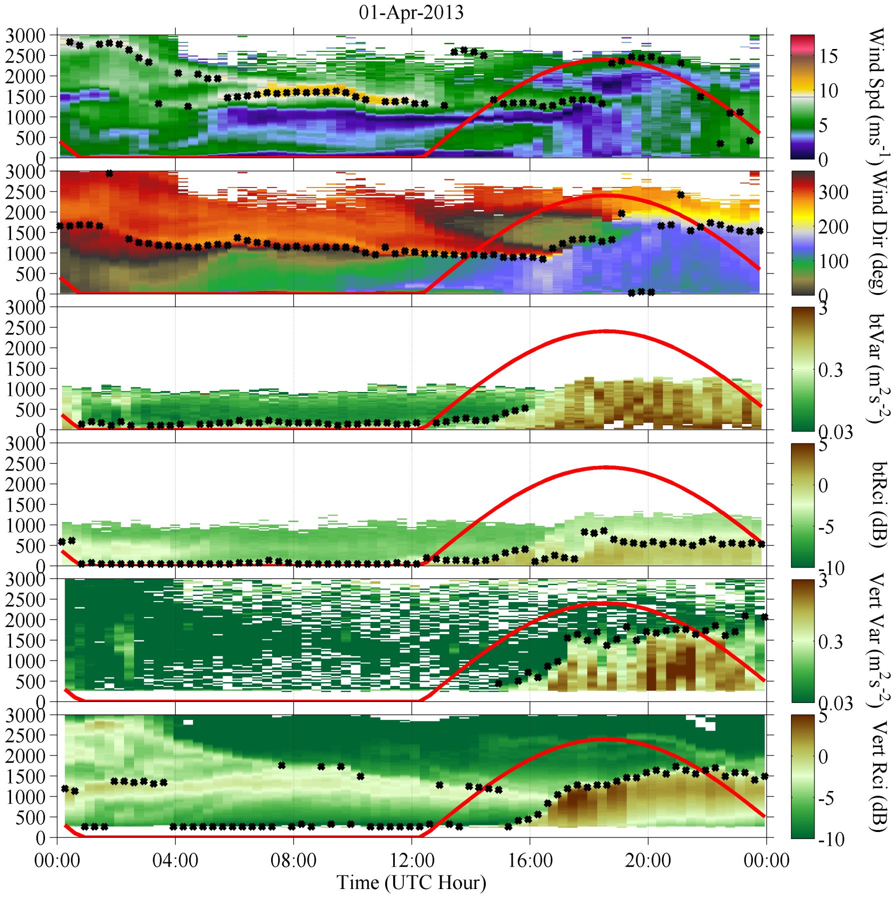


Figure 4.8: Results with *customized continuity test*. The table of settings is 4.4.

Estimator Parameter	$Z(\beta_{VS})$	$Z(\sigma_w)$	$Z(\beta_{BT})$	$Z(\sigma_h)$	$Z(W)$	$Z(W,\theta)$
Type of estimation	Haar wavelet transform	Peak-based thresholding	Haar wavelet transform	Peak-based thresholding	Peak in the profile	Haar wavelet transform
Threshold peaks?	yes					
Threshold value	0.1	0.5	0.1	0.2	3	0.75
Apply continuity test?	yes				no	yes
Δz	7	10	3	5		5
Δt	2	3	1	2		3
Chosen peak	lowest	In longest thread			strongest	longest

Table 4.4: Table of settings no. 4. Recap settings of the peak detection program for each estimator. The estimators are named according to table 3.2, page 20.

4.1.3 Cluster analysis

In this section, the results of the cluster analysis methods will be presented. The day for the test will be the same as for the assessment of peak detection method (01-Apr-2013 : few missing data, clear sky, fair diurnal cycle). The assessment of the cluster analysis is necessarily a bit different from the peak detection method. Indeed, the advantage of cluster analysis is to provide one estimation from many different data. While peak detection provides almost one estimation per type of data. The estimation will still be plotted as dots on a background of data, but now the dots are the same in every panel. And instead of looking at the peaks, we will look at the cluster map : a map of all the measurement points (time in x-axis, height in y-axis) with the number of the cluster the point belongs in shade of color.

The first paragraph will be devoted to present the classic clustering. As for peak detection methods, we will summarize the main settings (extended in the section 3.2.2) in a table : table 4.5.

Parameter	Value
K	2
Feature selection	$\begin{cases} \{\beta_{BT} \,,\, \sigma_h\} & \text{at night} \\ \{\beta_{VS} \,,\, \sigma_w \,,\, \beta_{BT} \,,\, \sigma_h\} & \text{at morning} \\ \{\beta_{VS} \,,\, \sigma_w\} & \text{at day} \\ \{\beta_{VS} \,,\, \sigma_w \,,\, \beta_{BT} \,,\, \sigma_h\} & \text{at evening} \end{cases}$

Table 4.5: Table of the settings for the cluster analysis method.

Classic clustering (p. 43) The results shown on the figure 4.9 come from the classic clustering, with $K = 2$ clusters. The features used vary depending on the time of the day. The choice of these features results from the observations and from the scan properties. Hence, during the night bowtie scan is used, during the day vertical staring is used, and both are used during the transition period. We don't use the wind information because it doesn't track the boundary layer shape under every condition[1].

On the figure 4.10 is shown the cluster map of the day. The color corresponds to the cluster the point belongs. For example the blue color corresponds to cluster 1, which is the "boundary layer cluster". On x-axis is the time index (one time index for each profile) and on y-axis is height in meters. Though we create only 2 clusters ("boundary layer" in blue and "free atmosphere" in red), there is a lot of white on the map. The white zone is excluded from the clustering domain because there is no data. It helps to visualize the time zone when different features are used. During the night, we use bowtie scan, which has a very low maximum range. This is visible by the upper limit of the color around 1000 meters. During the day, we use the vertical staring, which has a high minimum range. This is visible by the white stripe at the bottom (below 300 meters, roughly). During the morning, both are used, so the white zone is restricted where none of the two scan provide data.

During morning, the blue cluster ("boundary layer") appears several time in the profile. This appears when there are residual layers, or when the transition between boundary layer and free atmosphere is not very sharp. Here it is most likely due to the residual layer, we have seen very clearly in the previous figures. This is one limit of the method: if we want to have truly representative clusters, it might require more than two, especially at night. The relevant number of clusters is unknown *a priori*. Since the K-means algorithm needs an *a priori* number of clusters, it is impossible to use this algorithm to decide dynamically the number of clusters. Some variants of K-means are able to do it. Jain et al. [1999] evoke the existence of such algorithms (section 5.2.1, p. 16). The hierarchical clustering could be a good solution too, though it could be greedy in time. For this algorithm, the aim is only to find the top of the boundary layer, so we can make small compromises on the representativeness of the whole cluster, as long as the transition is well captured. The black line is the limit of the "boundary layer cluster" as seen by the algorithm. It corresponds to the dots on the figure 4.9.

[1] It could be useful in case of LLJ event, wind gradient or shear on the entrainment zone.

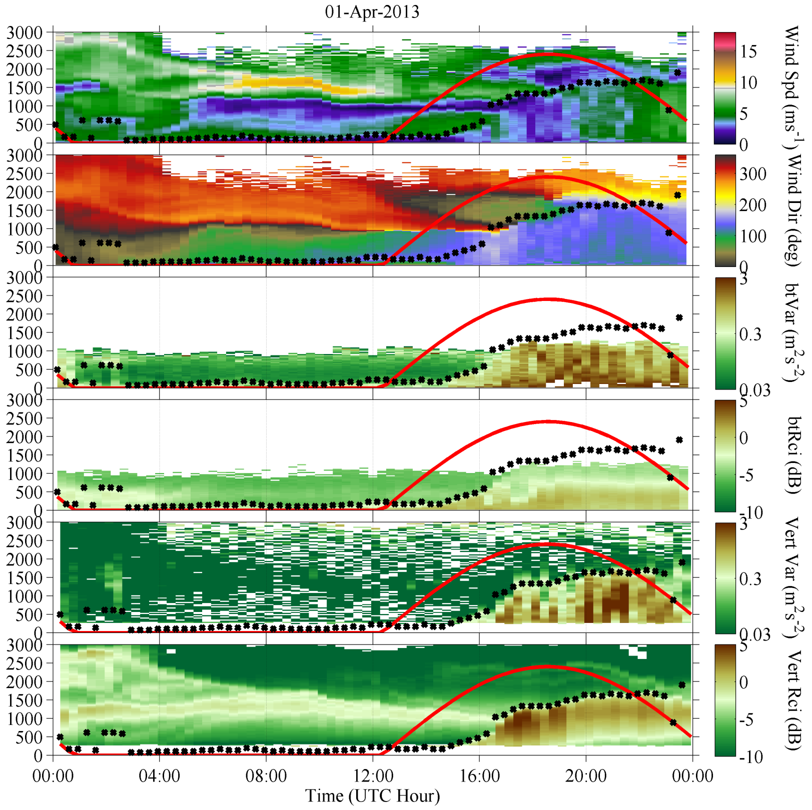


Figure 4.9: Results of cluster analysis (black dots) over the data (colors). The corresponding table of settings is 4.5.

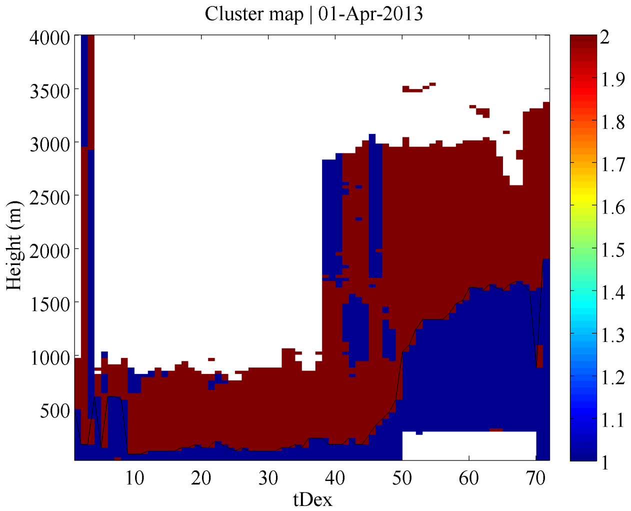


Figure 4.10: Cluster map. The blue zone corresponds to the BL cluster, the red zone corresponds to the free atmosphere. The black line is the detected BL top.
x-axis : time index, y-axis : height.

Fuzzy clustering (p. 45) Fuzzy clustering is very similar to the classic clustering. The only difference is that a pattern doesn't belong to only one cluster. It belongs partially to all clusters.

The advantage of the fuzzy clustering compared to classic clustering is that it could provides information about uncertainty. Classic clustering doesn't have the possibility to provide such an information because the membership is a logical. A pattern belongs to one and only one cluster, so the limit of a cluster is the pattern where the membership drops from to 1 to 0. This limit is always sharp, whatever the cluster are representative. With fuzzy clustering, the limit is not that sharp. It's even possible to find it by different ways, but the limit of a cluster is the pattern where the membership drops below 0.5. If the BL top is not well defined, the smoother transition gives theoretically a smoother decrease of the membership. The way the membership function varies is an indication of the uncertainty of the BL top. But because the profiles (and the membership as well) are discrete, exploiting the shape of the membership function is an issue.

The cluster map resulting from fuzzy clustering, with the same settings as in table 4.5 is displayed in figure 4.11. The blue color is still tracking the boundary layer air, and the red signalize the free atmosphere. But conversely to the previous cluster map (fig. 4.10), others colors are present. The color bar is the fuzzy cluster number $\tilde{k}$, define by the formula 4.1.

$$\tilde{k}(\mathbf{P}) = \sum_{k=1}^{K} k\, w_k(\mathbf{P}) \tag{4.1}$$

This is the fuzzy equivalent of the cluster number displayed on figure 4.10.

We can see on the fuzzy cluster map 4.11 the same elements as in the classic one : a shallow BL at night, the rise and the establishment of the convective BL later. The domain of clustering is limited in the same way, because we use the same data. But there are more details on this map than on the other. For example, the residual layer around 1600 meters in the morning is more visible here. The thin yellow thread between boundary layer and free atmosphere could be interpreted as the entrainment zone. It is thicker during the day than during the night.

Another difference is the presence of profiles completely yellow. That corresponds to a fuzzy cluster number close to 1.5 all along the profile. The method is not mature enough to be sure that the yellow profiles don't come from a crash of the iterative process in a local minimum. This question deserves further investigation. But it really comes from the data, it could mean that there isn't any truly representative cluster. That is to say the boundary layer top is ill-defined. But seeing that this yellow profiles occur during the day, where the transition is usually easier to catch, it is likely due to a problem in the convergence.

Fuzzy clustering seems to be a promising method to provide both a good estimation of BL and its uncertainty. The link between the membership and the data is still to be verified. But if the transition zone really gets thicker with the uncertainty, it will be a method of choice to evaluate the BLH.

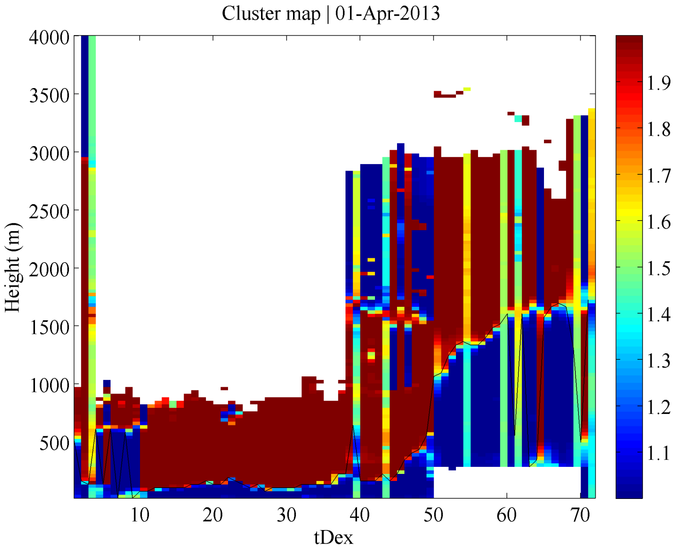


Figure 4.11: Cluster map with fuzzy clustering. The color bar represents the fuzzy cluster number (defined by formula 4.1). Blue colors still correspond to BL, and red correspond to free atmosphere. The shades of colors in between (yellow mostly) signalize a transition or an uncertainty. The fuzzifier value is $\phi = 2$.

4.2 Appraisal on the whole data set

4.2.1 Number of cluster validation

The number of cluster is an input parameter of the K-means algorithm. Unfortunately, the layering of the atmosphere may need more (or less) clusters than the given number to catch the BL properly. Too few clusters can lead to heterogeneous clusters. The clustering is less efficient in clarifying the structure in the profile because the clusters can't be interpreted. For example, this what happens for some morning profiles in the figure 4.10, page 43, when some blue patches are found out of the boundary layer. On the other hand, too many clusters will force the program to find limits where there is not. Several clusters can have the same interpretation. As a consequence, their limits will be variable from one run to another. Since it's not possible to dynamically choose the number of clusters with this version of K-means algorithm, the number have to be checked afterward.

The method used to check the number of clusters consists in checking the representativeness of the clusters for different number of clusters. There is no absolute manner to check the representativeness of a set of clusters. Some scores have been built to assess this representativeness. All of them have pros and cons, that is why two different scores will be tested. They are described by Pakhira et al. [2004] as widely used and efficient. They are also used by Toledo et al. [2013] as a criterion to choose the number of cluster.

The first score to be presented is the Dunn index Du. It is defined by the equation 4.2.

$$Du = \frac{\min\limits_{1\leqslant i\leqslant K}\left(\min\limits_{\substack{1\leqslant j\leqslant K\\ j\neq i}}\left[\delta(\mathbf{C}^i,\mathbf{C}^j)\right]\right)}{\max\limits_{1\leqslant k\leqslant K}\left(\max\limits_{1\leqslant \ell\leqslant n}\left[w_k(\mathbf{P}^\ell)\delta(\mathbf{C}^k,\mathbf{P}^\ell)\right]\right)} \tag{4.2}$$

The notation are the same as in section 3.2.2. This formula can be interpreted as the ratio of the smallest distance between different clusters over the width of the largest cluster. A well-separated set of compact cluster, will have a high distance inter-clusters $\min\limits_{1\leqslant i\leqslant K}\left(\min\limits_{\substack{1\leqslant j\leqslant K\\ j\neq i}}\left[\delta(\mathbf{C}^i,\mathbf{C}^j)\right]\right)$ and a small distance intra-cluster $\max\limits_{1\leqslant k\leqslant K}\left(\max\limits_{1\leqslant \ell\leqslant n}\left[w_k(\mathbf{P}^\ell)\delta(\mathbf{C}^k,\mathbf{P}^\ell)\right]\right)$. Hence, large values of the Dunn index are expected for well-separated compact clusters. The higher is the Dunn index, the better is our representativeness.

The second score is the Davies-Bouldin index *DB*, defined by the formula 4.3.

$$DB = \frac{1}{n}\sum_{i=1}^{n}\max_{\substack{1\leqslant j\leqslant K\\ j\neq i}}\left(\frac{\langle\delta(\mathbf{C}^i,\mathbf{P}^\cdot)\rangle + \langle\delta(\mathbf{C}^j,\mathbf{P}^\cdot)\rangle}{\delta(\mathbf{C}^i,\mathbf{C}^j)}\right) \tag{4.3}$$

$$\text{where } \left\langle\delta(\mathbf{C}^i,\mathbf{P}^\cdot)\right\rangle = \frac{\sum\limits_{\ell=1}^{n} w_i(\mathbf{P}^\ell)\delta(\mathbf{C}^i,\mathbf{P}^\ell)}{\sum\limits_{\ell=1}^{n} w_i(\mathbf{P}^\ell)}$$

$\langle\delta(\mathbf{C}^i,\mathbf{P}^\cdot)\rangle$ is the average point-to-seed distance in the cluster. It will be low is the cluster are compact and well-separated. On the other hand, the distance intra-cluster $\delta(\mathbf{C}^i,\mathbf{C}^j)$ will be large. In consequence, the Davies-Bouldin index will be small for homogeneous clusters, The smaller *DB* is, the better is our representativeness.

In the figure 4.12 are plotted the Davies-Bouldin index (upper panel) and the Dunn index (bottom panel) for different number of clusters, and for the four periods of the day. On the x-axis is the number of clusters (from 2 to 6). On the y-axis are the values of the index. The value itself doesn't matter, but the evolution does. The calculation have been done on all the profiles available, grouped according to the time of the day. Each profile provides a *DB* and a *Du* for all the number of clusters tested. What is plotted here is the average of all the profile, for all the number of cluster tested. The grouping along the time of the day ensure that the same features are used for the calculations. Whatever the period of the day, the number of clusters that provide the most representativeness is 2 (low *DB* and high *Du*). Hence the choice of $K = 2$ made in the section 3.2.2.

The limit of this evaluation is that it doesn't take into account the profile's particularities. Some profiles may need more than 2 clusters, but because the averaging smooth it. Since no solution has been implemented to allow a dynamic choice of K, this validation is the best that can be done.

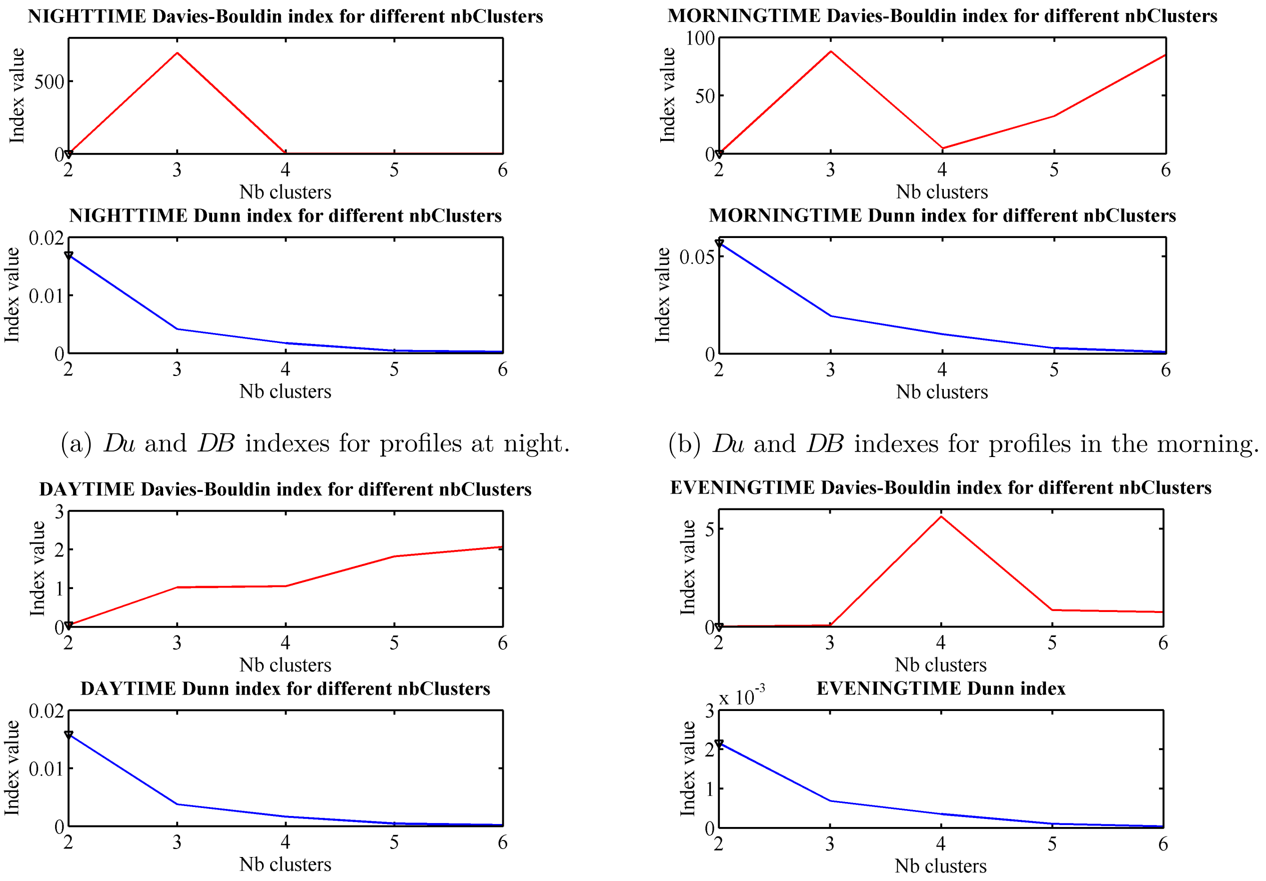
4.2.2 Visual examination

The difficulty in evaluating whether the method works or not is that there is nothing to compare with. The only way to say if the method gives a good result is to do visual examination. But the problem is then the quantity of profiles. Nevertheless, a visual examination has been done to give an idea of the percentage of success. To avoid the part of subjectivity that remains in visual examination, the assessment has been done twice independently, by two different persons. Since the BLH estimation will be applied profile by profile in the future field campaign, the estimators have been assessed without applying the continuity test, but the peaks are thresholded.

The method of examination is based on figures like 4.8. The result is superimposed on the data and a visual examination is possible. In order to keep some accuracy in this examination, each day is split into 4 hours chunks (signalized by the vertical dotted line). By comparing the dots to the data in background, a chunk is classified among three classes :

- "Successful" (S) is chosen when the estimation follow the BLH identified on the data.
- "Data missing" (D) is chosen when the chunk suffers from a lack of data, that makes the estimation unavailable or wrong or indeterminate.
- "Unsuccessful" (U) is chosen when the estimation is wrong though the data is good.

The visual examination is done day-by-day by counting the chunks in each of the 3 classes. It remains to fill in a table similar to table 4.6.



(a) *Du* and *DB* indexes for profiles at night.

(b) *Du* and *DB* indexes for profiles in the morning.

(c) *Du* and *DB* indexes for profiles at day.

(d) *Du* and *DB* indexes for profiles in the evening.

Figure 4.12: *Du* (bottom panel) and *DB* (top panel) indexes for different number of clusters, regrouped according to the time of the day. The black triangle signalize the best number of clusters (high *Du* and low *DB*).

Nevertheless, some cases need disambiguation. For example, the limitation of range are litigious. the maximum height of the bowtie scan is too small to catch the convective boundary layer, but that doesn't mean that the method is wrong. Instead of catching the boundary layer top, it may catch another layer reachable, and identifies it as the BLH. Same thing with the vertically staring data: the minimum range is too high to catch the shallow nocturnal layer. It may catch the residual layer instead, especially on the RCI profile. The membership of these cases to any of the classes above could be discuss. Since this is an instrumental bound, it's difficult to count it as "Unsuccessful", because it cannot be successful. When the boundary layer top is out of the range of a data, the estimator using this data is classified in "Data missing". As a consequence, the number of "Data missing" doesn't make any difference between the data missing at random (due to external factors like rain or clouds) and the data missing because of geometry (due to the scan itself, and so unavoidable and known).

Another case are the estimators based on wind speed and wind direction. The problem is that these methods will be accurate only if there is a certain wind structure in the boundary layer (low-level jet, wind shear). When there isn't such a structure, these estimators are unable to detect the BLH, though the data is reliable and the method works. In these case, the chunk will be classified in "Unsuccessful", because the meteorological conditions could also be responsible of the failure of the other estimators.

The results of the first visual examination are reported on the figure 4.13. For each estimator, the data set is represented as a vertical bar of height 1. The y-axis is the percentage of the data set. On the x-axis are the estimators. The "Successful" chunks (S) are the portion of the bar in blue, the "Data missing" ones (D) are in yellow, and the "Unsuccessful" (U) are in red.

The numerical values are in the table 4.7. Two columns are added to the raw data. The first one (4^{th}

Estimator Day	Estimator 1			Estimator 2			...
Day 1	# of S chunks	# of D chunks	# of U chunks	# of S chunks	# of D chunks	# of U chunks	...
Day 2	# of S chunks	# of D chunks	# of U chunks	# of S chunks	# of D chunks	# of U chunks	...
⋮	...						

Table 4.6: Example of table used to make the visual examination.

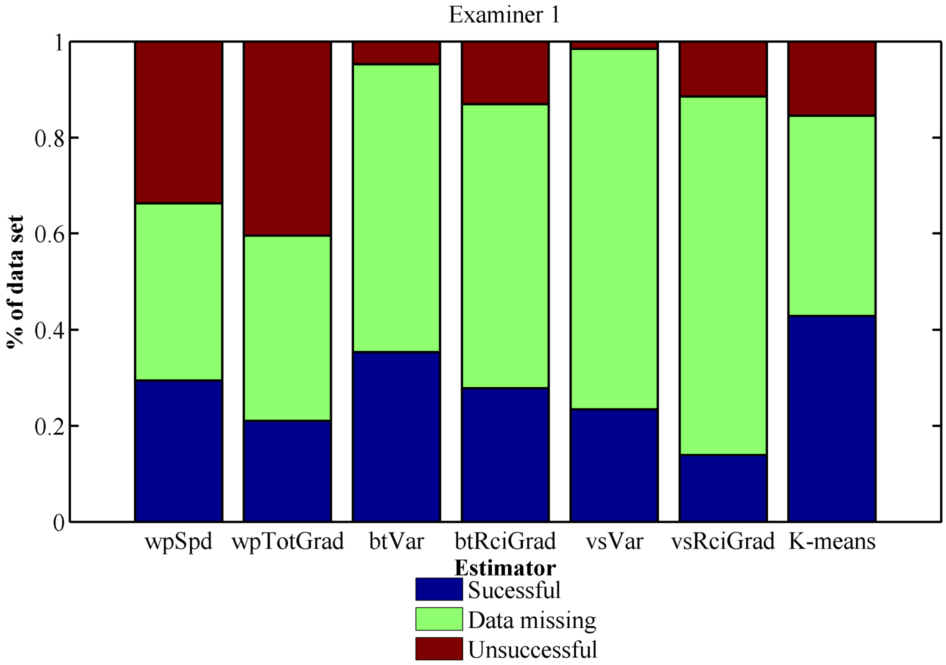


Figure 4.13: Results of visual examination by the first examinator. For each estimator (vertical bar), the data set is portioned in three classes: "Successful" (blue part), "Data missing" (yellow part) and "Unsuccessful" (red part).

column) is the *success rate*, defined by

$$s = \frac{S}{S+U}$$

It represents the percentage of success when the data is available. The second one (5$^{\text{th}}$ column) is the *availability rate*, defined by

$$a = \frac{S+U}{S+D+U}$$

It represents the fraction of the data set where the estimator doesn't suffer from a lack of data. In the "Data missing", both instrument limitations and external factors are counted. This is an issue for the scans like BT or VS, because their high amount of missing data is mostly due to the instrument limitations (geometry of the scans). The blue cells highlight the best estimator in each column.

The two first estimators are from the horizontal wind information (speed and direction). Both of them have a lot of "Unsuccessful" chunks (33.7% for $Z(W)$, 34.2% for $Z(W,\theta)$). However, they have the highest availability rate (respectively 0.63 and 0.61). The reason why they don't work often compared to the others is that they are more dependent on the meteorological conditions: if there is no structure such as low-level jet or wind shear, they cannot be accurate, even though the data is available. A fraction of this

Estimator	% of S	% of D	% of U	$\frac{S}{S+U}$	$\frac{S+U}{S+D+U}$
$Z(W)$	0.2937	0.369	0.3373	0.4654	0.631
$Z(W,\theta)$	0.2103	0.3849	0.4048	0.3419	0.6151
$Z(\sigma_h)$	0.3532	0.5992	0.0476	0.8812	0.4008
$Z(\beta_{BT})$	0.2778	0.5913	0.1310	0.6796	0.4087
$Z(\sigma_w)$	0.2341	0.75	0.0159	0.9375	0.25
$Z(\beta_{VS})$	0.1389	0.746	0.1151	0.5469	0.254
$Z_K(\beta,\sigma)$	0.4286	0.4167	0.1548	0.7347	0.5833

Table 4.7: Results of the visual examination by Examiner 1. S is for "Successful", D is for "Data missing", U is for "Unsuccessful". Blue cell highlight the best estimator in the column.

successful rate quantifies methods accuracy, another fraction quantifies how often horizontal structures can be used to define the boundary layer.

The next two estimators are from bowtie scan (horizontal velocity variance and range-corrected intensity). Different methods are applied on these data: for σ_h BLH is provided by peak-based thresholding, for β_{BT} BLH is given by Haar wavelet transform method. And the difference is sensible on the results. For $Z(\sigma_h)$, the "Unsuccessful" chunks are few (4.8%), but the number of chunks with enough data are also few. For $Z(\beta_{BT})$, there are more "Unsuccessful" chunks (13%), and a comparable number of chunks in need of data (59.9% for $Z(\sigma_h)$, 59.1% for $Z(\beta_{BT})$). The number of missing data is similar because they are from the same scan. Part of the chunks flagged as D because of the range limitation of bowtie scan. Almost every chunk during daytime is flagged D because the bowtie scan cannot reach the top of the boundary layer. The difference of successful chunks is due to the difference of method. These figures tends to prefer the peak-based thresholding method to the Haar wavelet transform method. Another noticeable difference between these two methods is the behavior when the BL is out of reach. The peak-based thresholding method won't give any answer, while the Haar wavelet transform will give a wrong answer (it will detect something else within the range of bowtie).

The last two peak-based estimators are from vertical staring (vertical velocity variance and range-corrected intensity). Same as for bowtie, these two estimators use different techniques. The conlusion is the same as for bowtie too. The peak-based method has very few unsuccessful chunks (1.6%) and a lot in need of data (75%). Same proportion of missing data for the Haar wavelet method (74.6%) but more unsuccessful chunks (11.5%). This agreement between the VS and BT results confirms the differences of the methods. The peak-based thresholding method seems to be a reliable method, when available.

The cluster analysis estimation $Z_K(\beta,\sigma)$ has a high ratio of successful chunks (42.9%), an acceptable (compared to the others) ratio of missing data (41.7%), but a significant part of unsuccessful chunks. It seems to be a good compromise between accuracy and availability. Its success ratio s is not the highest one (93.7% for $Z(\sigma_w)$), but is correct (73.4%). On the other hand, its availability rate is fairly higher than $Z(\sigma_w)$ (25% for $Z(\sigma_w)$, 58.3% for $Z_K(\beta,\sigma)$).

In order to ensure (or invalidate) these conclusions, another visual examination has been carried out by another examiner. The results of this visual examination are shown on the figure 4.14.

Some differences with the figure 4.13 are visible. All the estimators seems to have a little more unsuccessful chunks, and a little less with missing data. Only $Z(\beta_{BT})$ has less successful chunks. However, the comparison of the estimators is the same, excepted for $Z(\beta_{BT})$ which is degraded.

According to both visual examinations, the peak-based thresholding method seems to be the most secure method, but its weakness is the availability. The Haar wavelet transform method appears less accurate, and suffer from lack of data as well. The estimators from horizontal wind are the most available, but their success ratio is very low because the meteorological conditions needed for their accuracy are not systematic. The cluster analysis method seems to be a good compromise, according to a high percentage of successful estimation, tempered by a reasonable percentage of unsuccessful estimation and missing data.

In order to take both independent evaluation into account, the final figures of this assessment will be

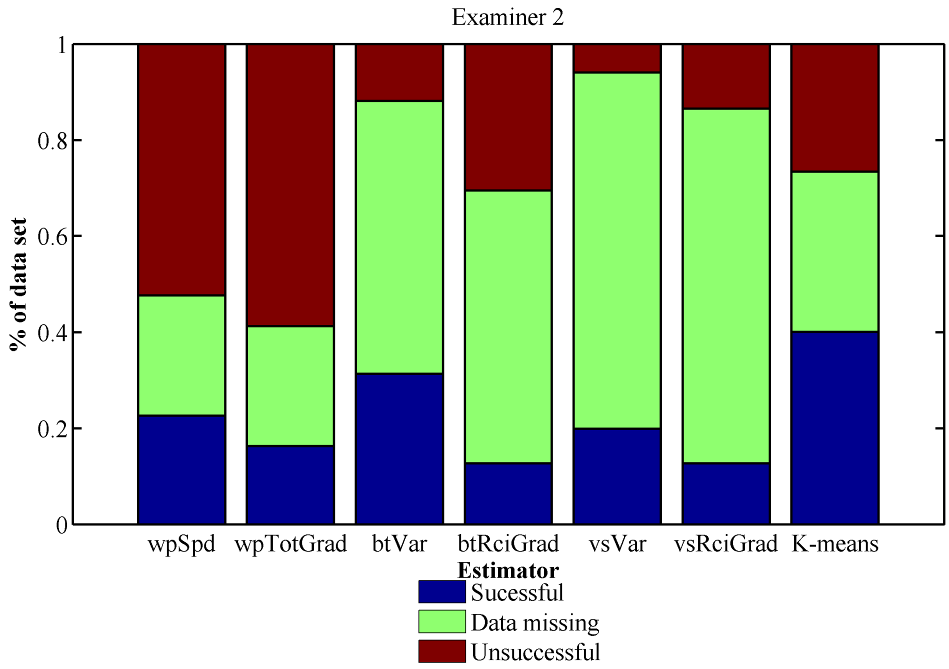


Figure 4.14: Results of visual examination. For each estimator (vertical bar), the data set is portioned in three classes : "Successful" (blue part), "Data missing" (yellow part) and "Unsuccessful" (red part).

the average of both evaluation. The final results are gathered in the table 4.8.

Estimator	% of S	% of D	% of U	$\frac{S}{S+U}$	$\frac{S+U}{S+D+U}$
$Z(W)$	0.25	0.31	0.43	0.37	0.69
$Z(W, \theta)$	0.18	0.31	0.49	0.27	0.68
$Z(\sigma_h)$	0.33	0.58	0.08	0.8	0.42
$Z(\beta_{BT})$	0.2	0.57	0.21	0.48	0.42
$Z(\sigma_w)$	0.21	0.75	0.03	0.85	0.25
$Z(\beta_{VS})$	0.13	0.74	0.13	0.52	0.26
$Z_K(\beta, \sigma)$	0.41	0.38	0.21	0.66	0.62

Table 4.8: Average results of the visual examinations. S is for "Successful", D is for "Data missing", U is for "Unsuccessful".

Conclusion

Several methods have been implemented to calculate boundary layer height from various data provided by Doppler lidar. The data used for this work were measured by HRDL, the NOAA's Doppler lidar. It provides vertical profiles of back-scattered intensity, horizontal and vertical velocity variance, horizontal wind speed and direction, every 20 minutes with a vertical resolution of 30 meters. These data are provided by three types of scans, with different ranges. The Haar wavelet transform method have been implemented to track gradients in these profiles. The profile is convoluted with a Haar function, and the boundary layer top is detected as a peak in the resulting profile. This method is used for the back-scattered profile. Another method involves peak detection directly on the profile and the peak are used to calculate a profile-dependent threshold. The boundary layer top is then the highest point above this threshold. This peak-based thresholding method has been applied to velocity variance profiles. To exploit horizontal wind information, two variants of the previous method have been tested. One aims to detect low-level jet by looking for peaks directly in the horizontal wind speed profiles. The other is built to detect shear, likely to appear at the top the capping inversion. The method applies the Haar wavelet method on both components of the horizontal wind, but the peaks are detected in the total gradient resulting from the Haar wavelet transforms. The last method to be developed is a cluster analysis method. The information are gathered into clusters using the K-means algorithm. Next, one of the cluster is identified as the boundary layer and the top is given by the border of the cluster. This method was built to integrate various forms of information, and uses back-scattered intensity and velocity variance from two scans.

These methods have been studied in a case study for one particular day. Next, they have been assessed by two independent visual examinations of the result over the original data. The difference of range among the scans has been taken into account for the appraisal. The most accurate method is the peak-thresholding method, with a success ratio around 80% (85% for VS.var, 80% for BT.var). The high amount of missing data (available 25% of the time of VS, 42% for BT) is mostly due to the geometry of the scans and cannot be avoided. A more realistic availability rate should distinguish the data missing at random (because of external factors like weather) and the data missing systematically (limitation of the instrument). The most available estimation come from wind information (68% of availability), but they suffer from a to big inaccuracy (only 37% of success ratio for $Z(W)$, and 27% for $Z(W, \theta)$). This inaccuracy can be explained by the frequency of the phenomena needed for this methods to be accurate. The Haar wavelet method is in the middle, with the same amount of missing data as peak-thresholding method (available 26% of the time of VS, 42% for BT), but lower success ration (52% for VS, 48% for BT). The cluster analysis is the best compromise between accuracy and availability. Its success ratio is of 66%, second highest after the peak-thresholding method, and its availability rate is of 62%, second highest after the wind information estimators.

Prospects

The closest next step to have a reliable estimate at any time of the day will be to increase the availability of the peak-thresholding method. This can be done by using the data provided by the bowtie scan at night and the vertical staring data during the day.

The fuzzy clustering method seems to be the natural next improvement of the K-means version implemented here. Since it has already been developed, the next step will be to investigate the convergence of the algorithm. The point is to ensure that the program doesn't get stuck in a local minimum of squared error. Once it is done, the link between the membership function and the boundary layer structure can be investigated to study if an uncertainty measurment by this technique is relevant.

An additional utilization of fuzzy clustering could be data imputation.

Bibliography

Brooks, I. M.
2003. Finding boundary layer top: Application of a wavelet covariance transform to lidar backscatter profiles. *Journal of Atmospheric & Oceanic Technology*, 20(8).

Browning, K. and R. Wexler
1968. The determination of kinematic properties of a wind field using doppler radar. *Journal of Applied Meteorology*, 7(1):105–113.

Cohn, S. A. and W. M. Angevine
2000. Boundary layer height and entrainment zone thickness measured by lidars and wind-profiling radars. *Journal of Applied Meteorology*, 39(8):1233–1247.

Gamage, N. and C. Hagelberg
1993. Detection and analysis of microfronts and associated coherent events using localized transforms. *Journal of the atmospheric sciences*, 50(5):750–756.

Grund, C.
1997. High resolution doppler lidar measurements of wind and turbulence. In *Advances in Atmospheric Remote Sensing with Lidar*, A. Ansmann, R. Neuber, P. Rairoux, and U. Wandinger, eds., Pp. 235–238. Springer Berlin Heidelberg.

Grund, C. J., R. M. Banta, J. L. George, J. N. Howell, M. J. Post, R. A. Richter, and A. M. Weickmann
2001. High-resolution doppler lidar for boundary layer and cloud research. *Journal of Atmospheric and Oceanic Technology*, 18(3):376–393.

Harvey, N. J., R. J. Hogan, and H. F. Dacre
2013. A method to diagnose boundary-layer type using doppler lidar. *Quarterly Journal of the Royal Meteorological Society*, 139(676):1681–1693.

Hayden, K., K. Anlauf, R. Hoff, J. Strapp, J. Bottenheim, H. Wiebe, F. Froude, J. Martin, D. Steyn, and I. McKendry
1997. The vertical chemical and meteorological structure of the boundary layer in the lower fraser valley during pacific'93. *Atmospheric Environment*, 31(14):2089–2105.

Jain, A. K., M. N. Murty, and P. J. Flynn
1999. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323.

Li, D., J. Deogun, W. Spaulding, and B. Shuart
2004. Towards missing data imputation: A study of fuzzy k-means clustering method. In *Rough Sets and Current Trends in Computing*, Pp. 573–579. Springer.

Melfi, S., J. Spinhirne, S. Chou, and S. Palm
1985. Lidar observations of vertically organized convection in the planetary boundary layer over the ocean. *Journal of climate and applied meteorology*, 24(8):806–821.

Menut, L., C. Flamant, J. Pelon, and P. H. Flamant
1999. Urban boundary-layer height determination from lidar measurements over the paris area. *Applied Optics*, 38(6):945–954.

Pakhira, M. K., S. Bandyopadhyay, and U. Maulik
2004. Validity index for crisp and fuzzy clusters. *Pattern recognition*, 37(3):487–501.

Pichugina, Y. L. and R. M. Banta
2010. Stable boundary layer depth from high-resolution measurements of the mean wind profile. *Journal of Applied Meteorology & Climatology*, 49(1).

Pollard, D. et al.
1981. Strong consistency of *k*-means clustering. *The Annals of Statistics*, 9(1):135–140.

Selim, S. Z. and M. A. Ismail
1984. K-means-type algorithms: a generalized convergence theorem and characterization of local optimality. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (1):81–87.

Senff, C., J. Bösenberg, G. Peters, and T. Schaberl
1996. Remote sensing of turbulent ozone fluxes and the ozone budget in the convective boundary layer with dial and radar-rass: A case study. *Contributions to atmospheric physics*, 69(1):161–176.

Stull, R. B.
1988. *An introduction to boundary layer meteorology*, volume 13. Springer.

Toledo, D., C. Córdoba-Jabonero, and M. Gil-Ojeda
2013. Cluster analysis: A new approach applied to lidar measurements for atmospheric boundary layer height estimation. *Journal of Atmospheric and Oceanic Technology*, 31(2013).

Tucker, Sara, C., A. Brewer, W., M. Banta, Robert, J. Senff, Christoph, P. Sandberg, Scott, C. Law, Daniel, M. Weikmann, Ann, and M. Hardesty, R.
2009. Doppler lidar estimation of mixing height using turbulence, shear and aerosol profiles. *Journal of Atmospheric and Oceanic Technology.*

List of Figures

List of Tables