第 43 卷 第 12 期/2023 年 6 月/光学学报 研究论文

光学学报

融合 K-means 和熵权法的高鲁棒性大气边界层高度估计方法

刘振兴[1,2,3]，常建华[1,2*]，李红旭[4]，孟园园[1]，周妹[1]，戴腾飞[1,2]

[1]南京信息工程大学电子与信息工程学院，江苏 南京 210044；

[2]南京信息工程大学江苏省大气环境与装备技术协同创新中心，江苏 南京 210044；

[3]泰州职业技术学院信息技术学院，江苏 泰州 225300；

[4]无锡学院电子信息工程学院，江苏 无锡 214105

摘要 针对常用激光雷达边界层高度估计方法在云层或悬浮气溶胶层等复杂大气结构下会产生误判的问题，提出一种融合 K-means 和熵权法的高鲁棒性大气边界层高度估计方法。选取美国大气辐射测量项目南部大平原站点的微脉冲激光雷达数据，将 K-means 算法和熵权法应用于多种条件下的边界层高度估计，从初始参数选取和距离计算两个方面提升基于聚类分析的边界层高度的估计性能。实验结果表明：与常用激光雷达边界层高度估计方法相比，所提方法具有较强的抗干扰能力，能更好地追踪复杂大气结构下的边界层高度日变化过程；在晴朗无云天气和复杂大气结构下，其边界层高度的估计值与无线电探空仪边界层高度的测量值基本一致，相关系数分别为 0.9718 和 0.9175。所提方法具有较高的鲁棒性，可以可靠地估计多种条件下的大气边界层高度。

关键词 遥感；激光雷达；大气边界层高度；复杂大气结构；聚类

中图分类号 P407.5 **文献标志码** A **DOI**：10.3788/AOS221534

1 引 言

大气边界层是对流层的最低层，该层直接受地表的影响。大气边界层高度（ABLH）是大气边界层的重要参数，其范围从几百米到数千米，对于分析边界层内热辐射传输过程、了解空气污染状况和制定污染控制策略等具有重要的作用[1-2]。探测边界层高度的仪器有多种，如系留气球[3]、微波辐射计[4]、无线电探空仪[5]和激光雷达[6]。其中，无线电探空仪通过寻找位温廓线的最大梯度位置或比湿廓线的最小梯度位置确定边界层高度，是比较可靠的一种边界层高度探测工具。然而，边界层高度受太阳辐射和其他因素的影响，在一天中不断演变，无线电探空测试有限的发射次数无法监测边界层高度的演变过程。激光雷达是一种主动遥感探测工具，具有较高的时空分辨率，并且可连续和自动测量边界层高度，其与传统的无线电探空仪相比具有显著的优势[7]。基于激光雷达数据估计边界层高度的方法主要有阈值法[8]、梯度法[9]、小波协方差变换法[10]和方差法[11]等。但是，这些方法仅适用于特定的气象条件，云和悬浮气溶胶层等的干扰会产生边界层高度的误判[12]。研究表明，不同天气状况的气溶胶变化存在很多不同之处，云和雾霾的存在会显著降低热通量，抑制边界层发展，从而减弱污染物的扩散，易导致重污染的发生[13]。此外，云是一种普遍的天气现象，覆盖了地球表面约三分之二的地区[14]。因此，在多云、悬浮气溶胶层、雾霾等复杂大气结构下估计 ABLH 是一项必要且具有挑战性的任务。

机器学习是一种强大的数据分析方法，广泛应用于光电探测[15-16]、目标识别[17]、机器视觉[18]等领域。本文将边界层高度的检测视为机器学习中的聚类问题，并探讨如何在复杂大气结构下进行边界层高度估计。Toledo 等[19]在不同大气条件下测试了 6 种常用的激光雷达边界层高度估计方法的鲁棒性，结果表明，在海陆微风和无尘条件下，6 种方法均取得了较好的测试结果，但是当残留层存在时，这些方法估计的边界层高度与无线电探空仪的测试值具有较大的误差。Thomas 等[20]描述了基于 K-means 和基于 AdaBoost 的两种激光雷达边界层高度检测方法，结果表明，整体上两种算法都具有较好的表现，但是 AdaBoost 受其训练数据的约束，而 K-means 估计的边界层高度受 K-means 初始

收稿日期：2022-07-26；**修回日期**：2022-09-13；**录用日期**：2022-10-14；**网络首发日期**：2022-10-24

基金项目：国家自然科学基金（61875089，62175114）、江苏高校“青蓝工程”资助项目（苏教师函〔2020〕10 号）、泰州市科技支撑计划社会发展项目（TSZ202132）、泰州职业技术学院院级重点科研项目（TZYKYZD-19-5）

通信作者：*jianhuachang@nuist.edu.cn

值、云层等的影响，具有较大的不确定性。

本文针对常用激光雷达边界层高度估计方法在复杂大气结构下存在的不足，提出一种融合 K-means 和熵权法（EK-means）的边界层高度估计方法，综合考虑聚类对象特征构建样本数据，通过对激光雷达后向散射信号梯度的分析确定 K-means 的初始值，采用带权重的欧氏距离对样本进行聚类，并基于聚类类别的特征对边界层高度进行估计。

2 数据来源与说明

本研究使用的地基微脉冲激光雷达和无线电探空数据来自美国大气辐射测量（ARM）项目南部大平原（SGP）站点（36°36′36″N，97°29′24″W）的中央设备（C1）[21]。选取 2002 年 1 月到 2004 年 5 月的数据作为研究对象，以下描述的设备为该时期所使用的设备。此外，本研究中还使用了与 SGP 观测点相邻的气溶胶自动观测网（AERONET）Cart 站点的数据，用于识别污染天气。

2.1 微脉冲激光雷达

SGP C1 的脉冲激光雷达发射波长为 523 nm，脉冲重复频率为 2.5 kHz，垂直分辨率为 30 m，时间分辨率为 30 s，最大探测距离可达 60 km。ARM 站点提供的微脉冲激光雷达数据为经过系统死区时间校正、背景噪声扣除、距离校正、重叠因子校正、后脉冲校正和能量归一化处理的相对后向散射信号（NRB），即

$$S_{\mathrm{NRB}}(r)=\frac{\left\{\left\{n(r)\times D\left[n(r)\right]\right\}-n_{\mathrm{ap}}(r)-n_{\mathrm{b}}\right\}r^{2}}{O_{\mathrm{c}}(r)E}=C\beta(r)T(r)^{2},\tag{1}$$

式中：r 为距离；$n(r)$ 为以每秒光子数表示的 r 处的返回信号；$D\left[n(r)\right]$ 为死区时间；$n_{\mathrm{ap}}(r)$ 为后向脉冲；n_{b} 为背景噪声；$O_{\mathrm{c}}(r)$ 为重叠因子；C 为系统校准常数；E 为发射的激光脉冲能量；β 为后向散射系数；T 为大气透过率；S_{NRB} 为 ARM 的增值数据产品，用于探测云和气溶胶，垂直分辨率为 90 m，时间分辨率为 1 min。此外，增值数据产品提供了云底高度、云顶高度和云顶数据衰减程度等信息，有利于进行实验验证。本文采用区间阈值技术[22]进行降噪处理，使用 4.37 km 以下的 NRB 数据，其中 120 m 以下的数据由于存在探测盲区，误差较大而被剔除。

2.2 无线电探空仪

为了评估基于激光雷达的边界层高度估计的准确性，将其与无线电探空仪数据进行对比。SGP 现场使用的无线电探空仪为 Vaisala RS90，通常每天发射 4 次，发射时间分别为 05：30、11：30、17：30 和 23：30 UTC，提供气压、温度、相对湿度等的垂直变化信息[23]。白天，地面热辐射会产生强烈的湍流和对流，使边界层内位温和气溶胶几乎呈均匀分布[5]，边界层变化明显。因此，本研究主要关注白天对流边界层高度的估计，当地时间为 UTC 时间减去 6。为了匹配激光雷达数据和无线电探空数据估计的 ABLH，激光雷达数据采用探空仪发射 10 min 内的平均值。

2.3 气溶胶自动观测网

AERONET 是全球地基气溶胶遥感观测网，该网络利用 CIMEL 自动太阳光度计作为基本观测仪器，为气溶胶研究和表征、卫星反演验证以及与其他数据库的协同作用提供了长期、连续和易于获取的气溶胶光学、微物理和辐射特性等公共领域数据库。AERONET 提供的数据产品包括气溶胶光学厚度（d_{AOD}）、反演产物和不同气溶胶状态下的可降水量的全球分布观测。气溶胶数据分为 3 个等级：Level 1.0、Level 1.5、Level 2.0。Córdoba-Jabonero 等[24]将 AERONET 提供的 Level 1.5 等级的气溶胶光学厚度和反演产物 Angstrom 指数（η_{AE}）作为沙尘天气的判断依据，指出沙尘天气发生时 $d_{\mathrm{AOD},500}>0.15$ 并且 $\eta_{\mathrm{AE},440/675}<0.5$。刘诏[25]指出，在晴朗无云的大气条件下，清洁天气的 $d_{\mathrm{AOD},500}<0.2$，阴霾天气的 $d_{\mathrm{AOD},500}>0.2$。基于此，本研究使用 AERONET 观测的 Cart 站点 $d_{\mathrm{AOD},500}$ 和 $\eta_{\mathrm{AE},440/675}$ 数据作为污染天气的判断依据，判断标准为污染天气 $d_{\mathrm{AOD},500}>0.2$ 并且 $\eta_{\mathrm{AE},440/675}<0.5$。

3 原理和方法

3.1 K-means

K-means 是一种经典的无监督机器学习算法，它根据样本到 k 个聚类中心的距离进行聚类。由于计算简单、效率高，该算法在许多应用中被广泛使用。基于 K-means 的 ABLH 估计方法的执行过程如下：

步骤 1　建立数据集 $\hat{\boldsymbol{X}}\in\mathbf{R}^{N\times F}$。$N$ 为数据点个数，F 为数据的维度，每个维度代表集群的一个特征。

步骤 2　归一化数据。不同维度上信息的数量级差别较大，为了确保各维度的影响效果，需对数据进行归一化处理。

$$X_{ij}=\frac{\hat{X}_{ij}-\mu_{\hat{X}}}{\sigma_{\hat{X}}},\tag{2}$$

式中：X_{ij} 为归一化后的数据；$i=1,\cdots,N$；$j=1,\cdots,F$；$\mu_{\hat{X}}$ 为均值；$\sigma_{\hat{X}}$ 为标准差。

步骤 3　选择聚类数量 k 和初始聚类中心 $\boldsymbol{C}$。通常，簇的数量 k 根据经验进行选取，初始中心为数据集内的随机位置。本文对聚类中心 k 和初始中心 $\boldsymbol{C}$ 的选取进行了研究，在后续 EK-means 中进行了详细描述。

步骤 4　根据欧氏距离对数据 $\boldsymbol{X}_i$ 聚类。

$$\boldsymbol{X}_i\in j,\qquad j=\mathrm{argmin}_j\, d(\boldsymbol{X}_i,\boldsymbol{C}_j),\tag{3}$$

式中：j 为聚类类别；$d(\boldsymbol{X}_i,\boldsymbol{C}_j)$ 为数据 $\boldsymbol{X}_i$ 与聚类中心 $\boldsymbol{C}_j$ 之间的欧氏距离。

步骤 5　使用簇中数据的平均值更新集群中心，定义为

$$C_j=\frac{1}{N_j}\sum_{X_i\in j}X_i, \tag{4}$$

式中：N_j为簇j的数据点数量。

重复步骤4和步骤5，直到达到最大迭代次数或聚类中心停止移动。在本研究中，由K-means确定的ABLH位于聚类强度从下到上第一次减弱的类别边界。

3.2 熵权法

熵权法是根据样本属性的变异程度来确定样本属性权重的一种方法。一般来说，样本属性变异程度越大，其信息熵越小，提供的信息量越多，在综合评价中的权重越大[26]。熵权法赋权的步骤如下：

步骤1 数据预处理。如前面所述，假设数据集$\hat{\boldsymbol{X}}\in\mathbf{R}^{N\times F}$，则数据预处理后有

$$Y_{ij}=\frac{\hat{X}_{ij}-\min\hat{X}_j}{\max\hat{X}_j-\min\hat{X}_j}。 \tag{5}$$

步骤2 计算信息熵，即

$$g_{ij}=Y_{ij}\Big/\sum_{i=1}^{N}Y_{ij}, \tag{6}$$

$$E_j=\begin{cases}-\dfrac{1}{\ln N}\displaystyle\sum_{i=1}^{N}g_{ij}\ln g_{ij}, & g_{ij}\neq 0\\ 0, & g_{ij}=0\end{cases}, \tag{7}$$

式中：E_j表示样本数据中第j个元素的信息熵。

步骤3 计算各属性权重：

$$W_j=\frac{1-E_j}{\sum_{j=1}^{F}(1-E_j)}。 \tag{8}$$

3.3 EK-means

常用K-means方法对聚类类别数和初始聚类中心很敏感，聚类结果随机性大。此外，常用于K-means的欧氏距离没有考虑多维样本中不同维度对聚类贡献的区别。针对这些问题，本文提出了EK-means方法。基于EK-means的ABLH估计流程如图1所示，该流程的执行过程如下：

步骤1 构建样本数据集。考虑到便捷性和普适性，本研究仅用激光雷达数据进行聚类分析来估计边界层高度。通过对簇的特征分析发现，在理想大气状态下，激光雷达后向散射信号的强度随探测高度升高而减小，但与云或其他悬浮气溶胶相互作用时，强度会急剧增加，并在其上部迅速减小，从而会产生较大的相对增长和梯度。本研究综合考虑气溶胶的时空分布特征，将高度r、归一化相对后向散射信号$S_{\mathrm{NRB}}(r)$和方差信号$S_{\mathrm{VAR}}\left[S_{\mathrm{NRB}}(r)\right]$作为样本的其中3个元素。此外，考虑到复杂大气条件，故将相对变化$r_{\mathrm{RI}}\left[S_{\mathrm{NRB}}(r)\right]$或梯度$G_{\mathrm{grd}}\left[S_{\mathrm{NRB}}(r)\right]=\mathrm{d}S_{\mathrm{NRB}}(r)/\mathrm{d}r$作为样本的第4个元素。

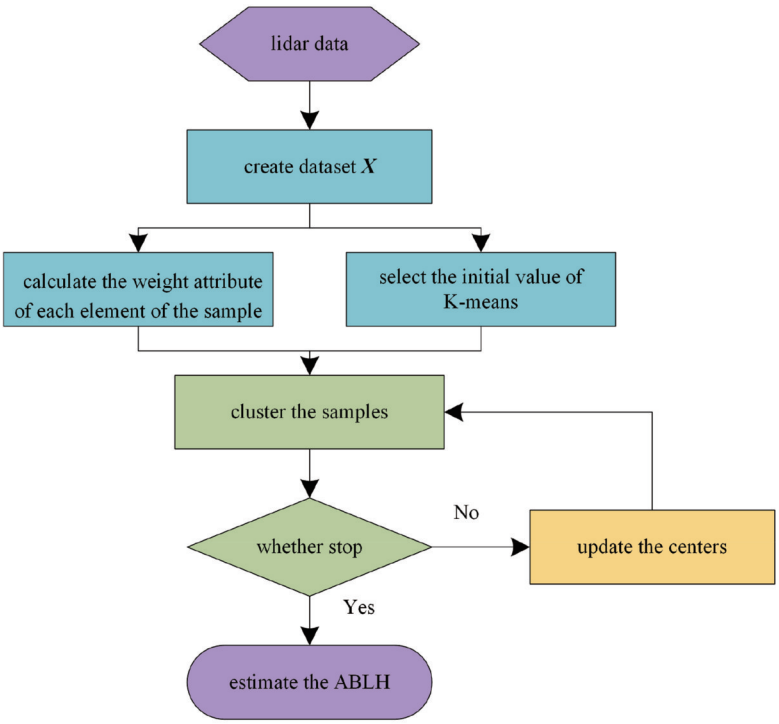

图1 基于EK-means的ABLH估计流程图

Fig. 1 Flowchart of ABLH estimation by EK-means from lidar data

图2为2004年3月22日17:32 UTC多云天气激光雷达后向散射信号特性图。从图2(a)可以看出，激光雷达回波信号整体上随高度的增加而减弱，在地面1.7～2 km范围内存在云层，回波信号强度在云层下方显著增大，而后在云层上方迅速衰减直至淹没。从图2(b)可以看出，在云层出现的位置产生了极强的正

负梯度信号，而云层上方由于信号衰减，产生的梯度信号非常弱。从图 2(c)可以看出，在云层出现的位置，产生了极强的相对增长信号，而云层上方虽然信号衰减，但产生的相对增长也非常强。研究表明，当相对增长大于 0.55 时，可判断为云层[21]，故通过观察相对增长廓线，可以猜测观测区域可能存在多个云层，而这与实际情况不一致。由此可知，在云层顶信号衰减较大的情况下不易将相对增长作为样本元素，否则会出现误判，产生较大的误差。因此，选取 r、$S_{\mathrm{NRB}}(r)$、$S_{\mathrm{VAR}}[S_{\mathrm{NRB}}(r)]$ 和 $\left|G_{\mathrm{grd}}[S_{\mathrm{NRB}}(r)]\right|$ 作为类别属性。

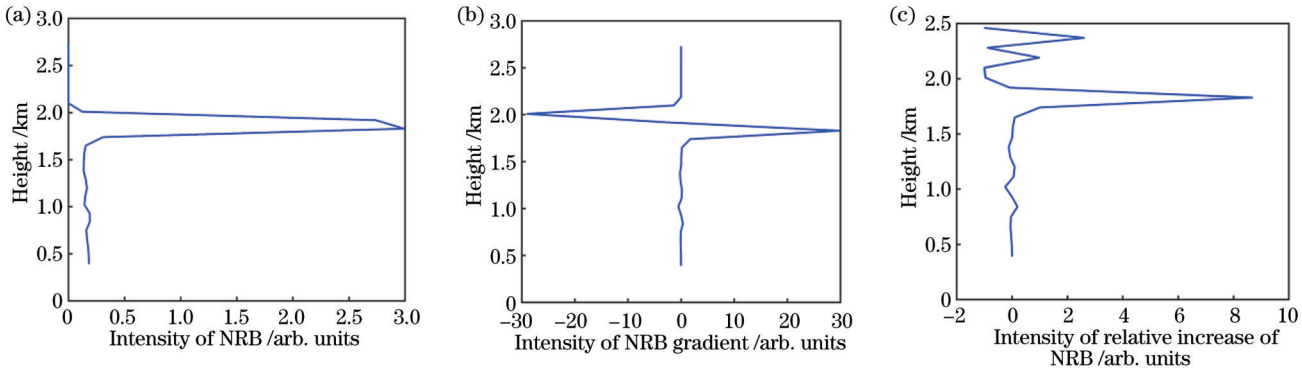

图 2 2004 年 3 月 22 日 17:32 UTC 多云天气 NRB 特性图。(a) NRB 垂直廓线；(b) NRB 梯度廓线；(c) NRB 相对增长廓线

Fig. 2 Characteristics of NRB on cloudy at 17:32 UTC, 22 March 2004. (a) Vertical distributions of NRB; (b) gradient of NRB; (c) relative increase of NRB

步骤 2 基于熵权法确定样本属性权重 $\boldsymbol{W}$。考虑到不同属性在基于欧氏距离的类别划分时贡献的大小，引入效用函数 $f(x)=x^2$。根据效用函数转换后的数据，按照式(5)～(8)计算样本各元素的权重属性 $\boldsymbol{W}=[W_1, W_2, W_3, W_4]$。

步骤 3 确定聚类类别数 k 和初始中心 $\boldsymbol{C}$。由前面的分析可知，当出现云层、悬浮气溶胶层时会产生强的信号变化，因此将梯度信号作为参照，选取聚类数 k。设想，在理想大气状态下，激光雷达回波信号强度随高度上升而减弱，整个观测范围内信号的梯度值为负值，此时回波信号经过间隔阈值处理后，同向区间数为 1，至少需要聚类为 2 类才能将边界层检测出来。当有云层存在时，回波信号在云层下方的强度随高度上升而增强，在云层上方的强度变化则相反，故会产生 3 个同向区间，如图 3 所示。

图 4 所示为来自图 2 的数据在不同簇数下的 K-means 聚类效果。由无线电探空数据可知，此时边界层在云层下方。由图 4 可知：当 $k=3$ 时，聚类为边界层、云层和自由对流层，此时易将云底高度估计为边界层高度，从而产生较大的误差；当 $k=4$ 时，聚类为边界层、云中间层、云边缘层和自由对流层，此时可以估计边界层高度；当 $k=5$ 时，聚类为边界层、自由对流层、云边缘层、云中间层和云上衰减层，此时可以估计边界层高度，与 $k=4$ 相比边界层高度有所下降；当 $k=6$ 时，对云层进行进一步分割，其增加了计算量，但对边界层的估计帮助不大。因此，通过计算梯度 $G_{\mathrm{grd}}[S_{\mathrm{NRB}}(r)]$，然后采用间隔阈值处理获取同向间隔数 n，可得聚类类别数 $k=n+1$ 或 $k=n+2$。根据实验可知，当云上的信号完全衰减时取 $k=n+2$，否则取 $k=n+1$。聚类初始中心选取同向区间的最大信号强度位置，其中第一个负值区间均匀选取两个中心，并采用 davis-bouldin 指数[20]进行微调。

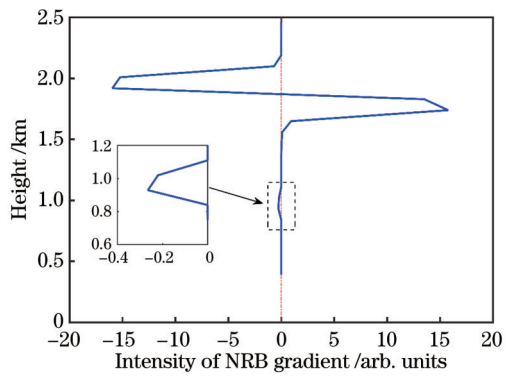

图 3 经区间阈值技术处理后的 NRB 梯度信号廓线

Fig. 3 Profile of NRB gradient after processed by interval thresholding technique

步骤 4 计算样本到中心的距离，进行聚类。

$$v(\boldsymbol{X}_i, \boldsymbol{C}_j)=\sqrt{\sum_{t=1}^{4} W_t(X_{it}-C_{jt})^2}, \tag{9}$$

$$\boldsymbol{X}_i \in j, \quad l=\mathrm{argmin}_j\, v(\boldsymbol{X}_i, \boldsymbol{C}_j), \tag{10}$$

式中：$l=1, \cdots, k$，表示聚类类别；$v(\boldsymbol{X}_i, \boldsymbol{C}_j)$ 为所提出的样本 $\boldsymbol{X}_i$ 与聚类中心 $\boldsymbol{C}_j$ 之间的距离。

步骤 5 参照式(4)更新中心，返回上一步重新计算各样本到中心的距离，并进行聚类，直至中心更新停止或达到最大迭代次数。

步骤 6 确定边界层高度。一般情况下，ABLH 位于聚类强度从下到上第一次下降的类别边界。

4 实验和讨论

对常用 K-means 方法进行改进，采用本文方法确

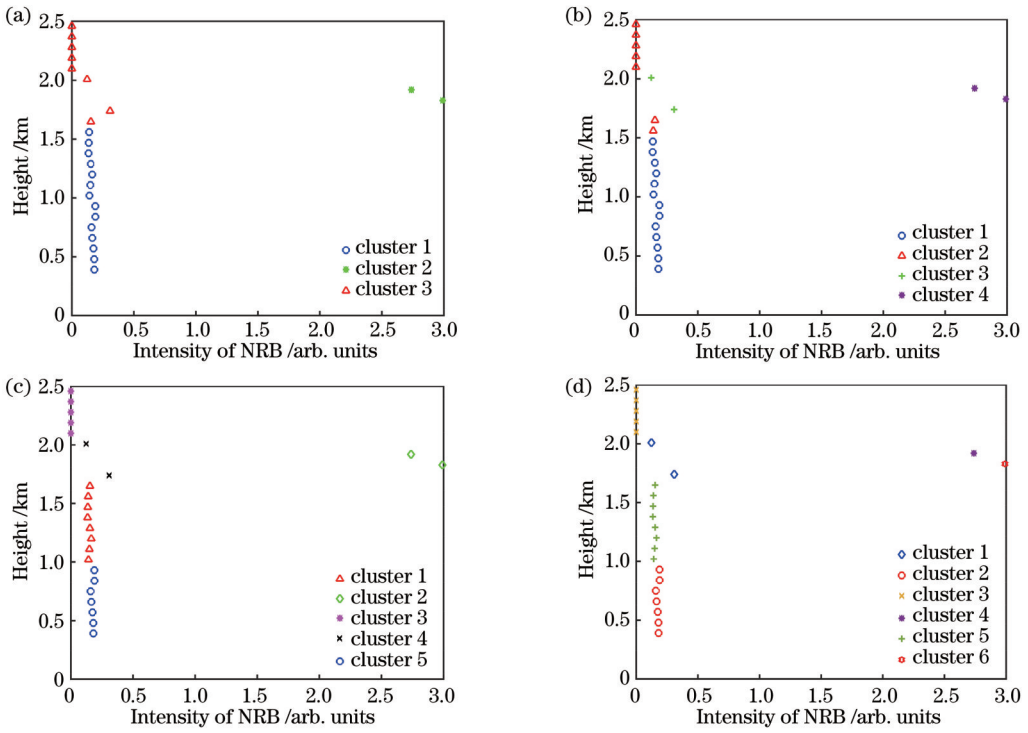

图 4　不同 k 值下的聚类效果。(a) k=3;(b) k=4;(c) k=5;(d) k=6

Fig. 4　Clustering effects under different clusters to profile of NRB. (a) k =3; (b) k =4; (c) k =5; (d) k =6

定初始值。将本文方法——EK-means 方法、改进后的 K-means 方法、小波协方差变换法和梯度法等基于激光雷达数据的方法进行边界层高度估计实验，分别选取晴朗无云天气、污染天气和多云天气等 3 种典型天气的对流边界层日变化过程进行分析，并将晴朗无云天气和多云或悬浮气溶胶层结构下激光雷达法估计的边界层高度与无线电探空仪测试结果进行比较。

4.1　激光雷达法估计的边界层高度日变化过程对比分析

图 5 所示为 2004 年 3 月 31 日晴朗无云天气下利用不同方法确定的边界层高度，横坐标表示 UTC 时间，纵坐标表示地表高度，背景信号为 NRB，GM 表示梯度法，WM 表示小波协方差变换法，RS 表示无线电探空仪，当天无线电探空仪发射的时刻为 17:29 和 23:30 UTC。从图 5 可以看出，当天边界层结构明显，在地表面 0.5～0.9 km 范围内有明显的信号突变，4 种激光雷达方法都可以很好地追踪边界层高度的日变化过程。在 16:30 UTC 之前，EK-means 和 K-means 方法估计的边界层高度略大于梯度法和小波协方差变换法的估计值；16:40—18:40 UTC 时间段，利用 EK-means 和 K-means 方法估计的边界层高度略低于梯度法和小波协方差变换法的估计值；其他时段，4 种方法估计的边界层高度有略微的差别，但是整体上这些方法估计的边界层高度差异不大，表现出很好的一致性。

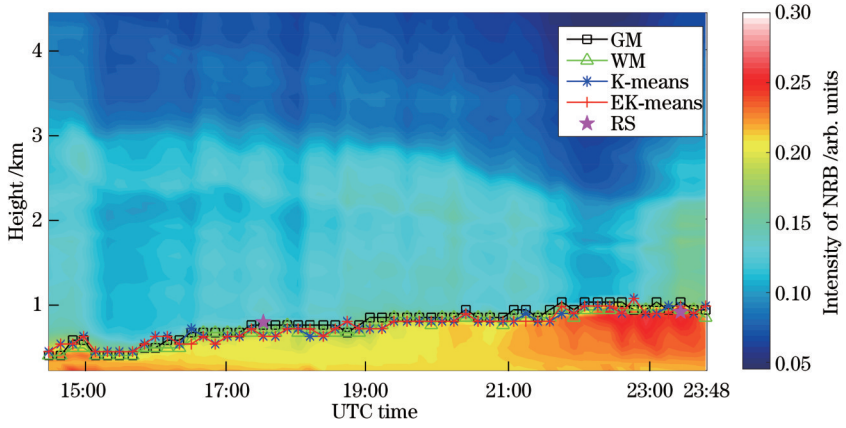

图 5　2004 年 3 月 31 日晴朗无云天气下的各种方法估计的 ABLH

Fig. 5　ABLH estimated by different methods on clear sky conditions on 31 March 2004

图 6 所示为 2002 年 5 月 17 日污染天气下利用不同方法确定的边界层高度，当天 17：29 UTC 时刻的无线电探空数据为无效值。根据 AERONET Cart_Site 站点数据，20：40—23：50 UTC 时段 $d_{\mathrm{AOD},500}>0.2$ 且 $\eta_{\mathrm{AE},440/675}<0.5$，故该时段为污染天气。由图 6 可知，15：00—20：00 UTC 时段地表面 0.5～1 km 高度范围内有云层存在，而在云层下方无强的信号衰减，故边界层高度在云层上方或与云层耦合。此时段，4 种方法皆将云顶处高度估计为边界层高度。20：10—23：50 UTC 时段，在地表面 1.4 km 高度处有较强的气溶胶信号存在，该位置附近气溶胶浓度值较高，与 AERONET 数据信息基本一致，可判断为高浓度的污染气溶胶层，此时 4 种方法将高浓度气溶胶层顶端高度估计为边界层高度。23：30 UTC 时刻 4 种方法的边界层高度估计值与无线电探空仪测量值基本一致。

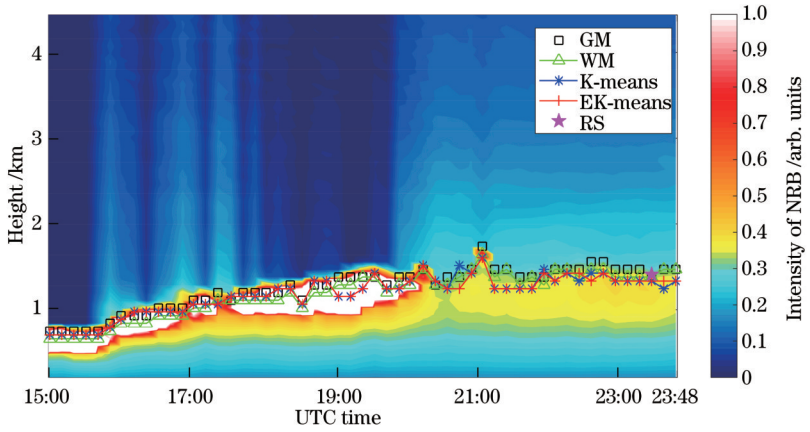

图 6　2002 年 5 月 17 日污染天气下不同方法估计的 ABLH

Fig. 6　ABLH estimated by different methods on polluted conditions on 17 May 2002

图 7 为 2004 年 3 月 22 日多云天气下利用不同方法估计的 ABLH 日变化示意图。由图 7 可知，当日气溶胶结构复杂，既存在云层，又存在悬浮气溶胶层。无线电探空仪发射时间为 17：29 和 23：30 UTC。从图 7 可以看出，观测的开始阶段，地面 1 km 和 1.3 km 处存在悬浮气溶胶层，梯度法和小波协方差变换法皆将悬浮气溶胶层顶高度估计为 ABLH，而 K-means 和 EK-means 能可靠地识别近地面的边界层。15：30—21：00 UTC 时段，地面 2 km 附近存在云层；21：00—24：00 UTC 时段，地面 1.6 km 处存在云层。梯度法和小波协方差变换法始终将气溶胶浓度变化最强烈的区域即云层上方的高度估计为 ABLH。然而，由图 7 可知，15：30—22：40 UTC 时段，云层下方有清晰可见的边界层，故利用梯度法和小波协方差变换法估计的边界层高度存在较大的误差。该时段，利用 K-means 和 EK-means 方法均能较好地捕捉到云层下方的边界层高度位置，其中 EK-means 的效果最好，能可靠识别并捕捉 ABLH 的变化过程。在 17：29 UTC，梯度法和小波协方差变换法估计的 ABLH 远远高于无线电探空仪的 ABLH 估计值，而 K-means 和 EK-means 测定的 ABLH 与无线电探空仪测定的 ABLH 具有较高的吻合度。23：00—23：20 UTC 时段，云层下方的气溶胶浓度变化不显著，4 种方法皆将云层顶高度当作 ABLH。23：30—23：50 UTC 时段，在地面 1 km 处存

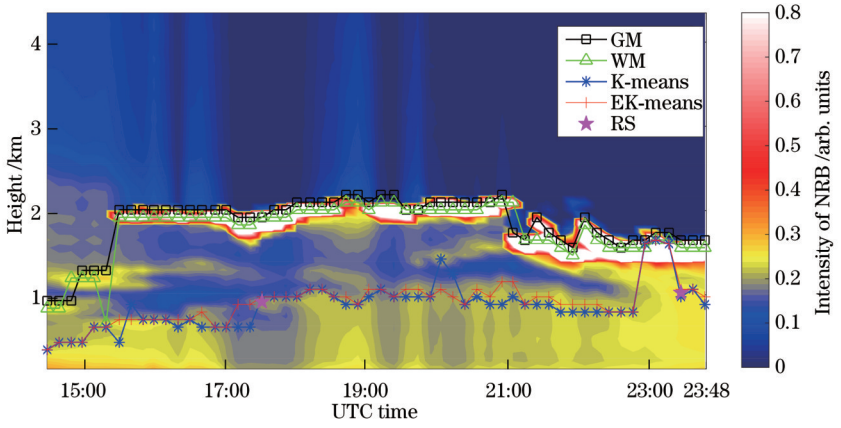

图 7　2004 年 3 月 22 日多云天气下不同方法估计的 ABLH

Fig. 7　ABLH estimated by different methods on cloudy conditions on 22 March 2004

在明显的气溶胶浓度突变，K-means和EK-means方法皆将该位置当作ABLH，其估计值与无线电探空仪测量值一致，梯度法和小波协方差变换法将云层顶的高度当作ABLH。

4.2 激光雷达法与无线电探空仪法测量的ABLH对比分析

图8所示为晴朗无云天气下基于激光雷达的方法估计的边界层高度与无线电探空仪测量的边界层高度的结果比较，其中R为相关系数，N_o为例子数量。本研究选取观测站点2003年1月至2004年5月46例边界层清晰的数据进行对比分析。从结果可以看出，在晴朗无云天气，利用4种激光雷达方法估计的边界层高度均匀分布在对角线$y=x$两侧，其边界层高度估计值与无线电探空仪的测量值具有很高的相关性，相关系数皆大于0.95，其中本文方法估计的边界层高度的相关系数最大，为0.9718。

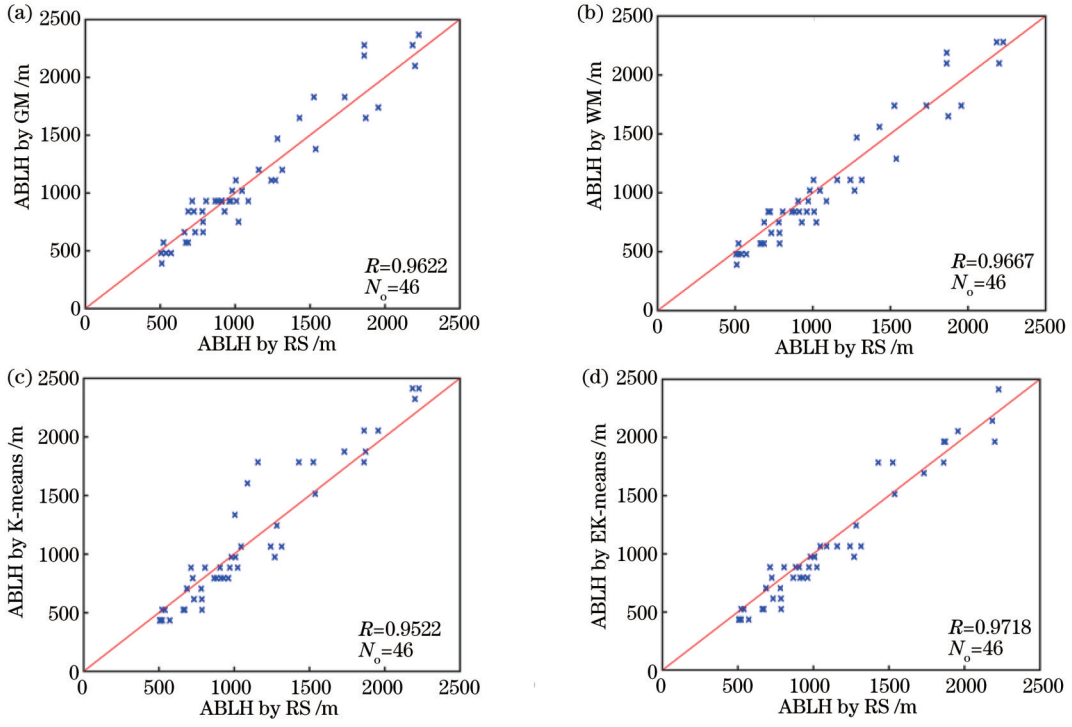

图8 晴朗无云天气激光雷达法估计的边界层高度与无线电探空仪测量值的比较。(a) GM和RS；(b) WM和RS；(c) K-means和RS；(d) EK-means和RS

Fig. 8 Comparisons between ABLH results determined by lidar-based methods and radiosonde on clear sky. (a) GM and RS; (b) WM and RS; (c) K-means and RS; (d) EK-means and RS

图9所示为多云天气或多层气溶胶结构下基于激光雷达方法估计的边界层高度与无线电探空仪的边界层高度测量结果对比。本研究选择了43例边界层清晰的数据进行比较。从图9(a)、(b)可以看出，当有云层或悬浮气溶胶层存在时，这些强信号会对梯度法和小波协方差变换法估计边界层高度产生严重的干扰，此时这两种方法估计的边界层高度与无线电探空仪测量的边界层高度具有较大的差异，前两者的估计值远远大于后者，其与无线电探空仪测量值相关系数分别为0.4247和0.4453。从图9(c)、(d)可以看出，利用K-means和EK-means方法估计的边界层高度接近无线电探空仪的测量值，其相关系数分别为0.7986和0.9175，可见本文方法显著提高了边界层高度估计的可靠性。

表1所示为与图8和图9对应的两种典型气象条件下4种激光雷达方法估计的边界层高度与无线电探空仪测量值的对比，采用相关系数R、绝对误差均值(MAE)和绝对误差中位数(MDAE)表征算法性能。从表1可以看出，在晴朗无云天气，利用这4种激光雷达方法估计的边界层高度与无线电探空仪测量的边界层高度具有较大的相关性，绝对误差均值和绝对误差中位数皆较小，其中本文方法性能最优，相关系数为0.9718，绝对误差均值为0.1059 km，绝对误差中位数为0.0869 km。在多云天气或悬浮气溶胶层结构下，利用梯度法和小波协方差法估计的边界层高度与无线电探空仪测量值相比，相关系数非常小，绝对误差均值和绝对误差中位数皆较大，可见强信号的存在严重干扰了梯度法和小波协方差法对边界层高度的估计，产生较大的误差；K-means和EK-means方法由于采用多个特征对观测对象进行聚类，具有较强的抗干扰性能，其中所提出的EK-means方法性能最优，相关系数为0.9175，绝对误差均值为0.1317 km，绝对误差中位数为0.1155 km。

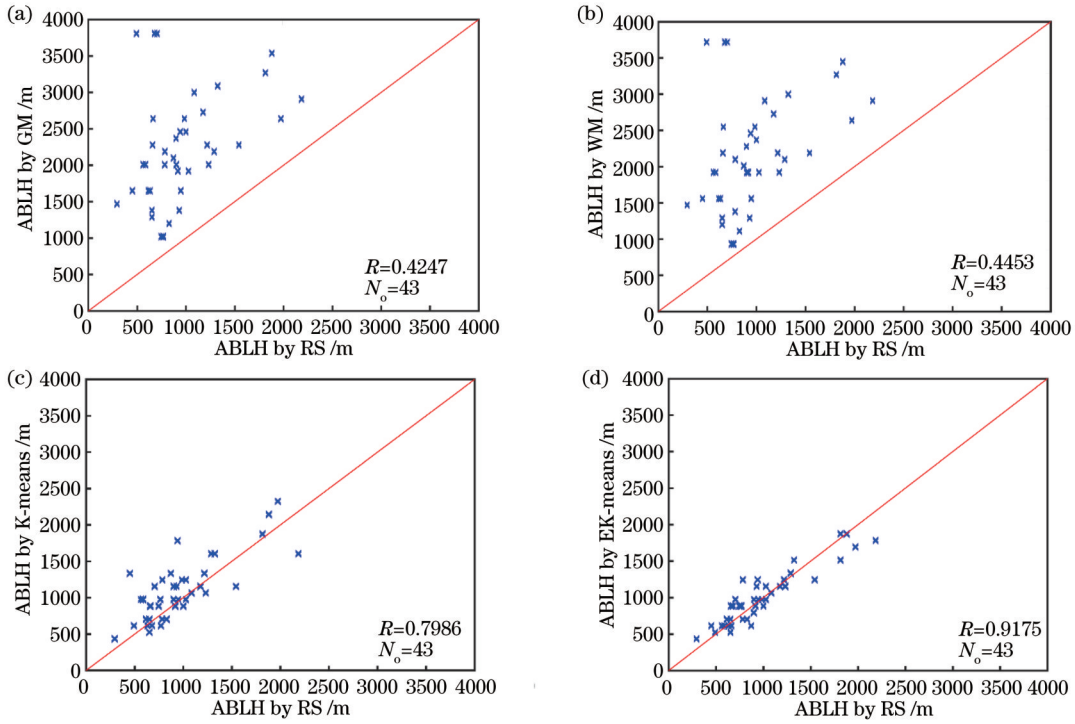

图 9 多云或悬浮气溶胶层结构下激光雷达法估计的边界层高度与无线电探空仪测量值的比较。(a) GM 和 RS;(b) WM 和 RS;(c) K-means 和 RS;(d) EK-means 和 RS

Fig. 9 Comparisons between ABLH results determined by lidar-based methods and radiosonde on cloudy or structure of the suspended aerosol layer. (a) GM and RS; (b) WM and RS; (c) K-means and RS; (d) EK-means and RS

表 1 基于激光雷达的方法估计的边界层高度与无线电探空仪测量的边界层高度的比较

Table 1 Comparison of ABLH determined by lidar-based methods and radiosonde

Atmospheric condition	Method	R	MAE / km	MDAE / km
Clear sky	GM	0.9622	0.1183	0.1018
	WM	0.9667	0.1182	0.1033
	K-means	0.9522	0.1404	0.1184
	EK-means	**0.9718**	**0.1059**	**0.0869**
Cloudy	GM	0.4247	1.2573	1.1980
	WM	0.4453	1.1757	1.1080
	K-means	0.7986	0.2246	0.1544
	EK-means	**0.9175**	**0.1317**	**0.1155**

5 结 论

边界层高度的可靠估计对于大气污染防治、天气预报和气候变化研究等具有重要作用。本文提出一种融合 K-means 和熵权法的高鲁棒性 ABLH 估计方法，通过对激光雷达后向散射梯度信号进行分析确定 K-means 初始参数，采用熵权法计算样本各特征的权重属性，改进了基于距离聚类的边界层高度估计方法。实验结果表明，相比于梯度法和小波协方差变换法等常用的激光雷达边界层高度估计法，所提方法在晴朗无云天气、污染天气和多云或悬浮气溶胶层结构等条件下都能更好地追踪边界层高度的日变化过程；在晴朗无云天气和多云或悬浮气溶胶层结构等条件下，其估计的边界层高度与无线电探空仪边界层高度测量值具有更好的一致性，相关系数更高，绝对误差均值更小。但当云层位于边界层内时，受限于探测方式，所提方法的优越性未能得到体现，后续工作将采用星载激光雷达和地基激光雷达联合探测的方式开展研究。

参 考 文 献

[1] 于思琪, 刘东, 徐继伟, 等. 基于激光雷达探测的金华、合肥和兰州大气边界层高度及其统计分析[J]. 光学学报, 2021, 41(24): 1422002.
Yu S Q, Liu D, Xu J W, et al. Statistics and analysis of planetary boundary layer height retrieved by lidar over Jinhua, Hefei, and Lanzhou[J]. Acta Optica Sinica, 2021, 41(24): 1422002.

[2] Shi Y, Hu F, Fan G Q, et al. Multiple technical observations of the atmospheric boundary layer structure of a red-alert haze episode in Beijing[J]. Atmospheric Measurement Techniques, 2019, 12(9): 4887-4901.

[3] Liu D Y, Yan W L, Kang Z M, et al. Boundary-layer features and regional transport process of an extreme haze pollution event in Nanjing, China[J]. Atmospheric Pollution Research, 2018, 9(6): 1088-1099.

[4] Min J S, Park M S, Chae J H, et al. Integrated System for Atmospheric Boundary Layer Height Estimation (ISABLE) using a ceilometer and microwave radiometer[J]. Atmospheric Measurement Techniques, 2020, 13(12): 6965-6987.

[5] Liu S Y, Liang X Z. Observed diurnal cycle climatology of planetary boundary layer height[J]. Journal of Climate, 2010, 23(21): 5790-5809.

[6] 项衍, 张天舒, 刘建国, 等. 基于激光雷达对WRF模式模拟边界层高度的评估[J]. 中国激光, 2019, 46(1): 0110002.
Xiang Y, Zhang T S, Liu J G, et al. Evaluation of boundary layer height simulated by WRF mode based on lidar[J]. Chinese Journal of Lasers, 2019, 46(1): 0110002.

[7] Caicedo V, Rappenglück B, Lefer B, et al. Comparison of aerosol lidar retrieval methods for boundary layer height detection using ceilometer aerosol backscatter data[J]. Atmospheric Measurement Techniques, 2017, 10(4): 1609-1622.

[8] Melfi S H, Spinhirne J D, Chou S H, et al. Lidar observations of vertically organized convection in the planetary boundary layer over the ocean[J]. Journal of Climate and Applied Meteorology, 1985, 24(8): 806-821.

[9] 刘娜娜, 罗涛, 韩亚娟, 等. 台风外围环流对沿海地区大气边界层结构的影响研究[J]. 光学学报, 2021, 41(19): 1901004.
Liu N N, Luo T, Han Y J, et al. Influence of typhoon peripheral circulation on atmospheric boundary layer structure in coastal areas[J]. Acta Optica Sinica, 2021, 41(19): 1901004.

[10] 于思琪, 刘东, 徐继伟, 等. 激光雷达反演大气边界层高度的优化方法[J]. 光学学报, 2021, 41(7): 0728002.
Yu S Q, Liu D, Xu J W, et al. Optimization method for planetary boundary layer height retrieval by lidar[J]. Acta Optica Sinica, 2021, 41(7): 0728002.

[11] Wang F T, Yang T, Wang Z F, et al. A comprehensive evaluation of planetary boundary layer height retrieval techniques using lidar data under different pollution scenarios[J]. Atmospheric Research, 2021, 253: 105483.

[12] Kotthaus S, Halios C H, Barlow J F, et al. Volume for pollution dispersion: London's atmospheric boundary layer during ClearfLo observed with two ground-based lidar types[J]. Atmospheric Environment, 2018, 190: 401-414.

[13] Quan J N, Gao Y, Zhang Q, et al. Evolution of planetary boundary layer under different weather conditions, and its impact on aerosol concentrations[J]. Particuology, 2013, 11(1): 34-40.

[14] 孟园园, 常建华, 陈思成, 等. 基于双向重构后向散射信号的微脉冲激光雷达云层检测算法[J]. 光学学报, 2022, 42(24): 2428003.
Meng Y Y, Chang J H, Chen S C, et al. Cloud detection algorithm of micro-pulse lidar based on bidirectional reconstruction of backscatter signal[J]. Acta Optica Sinica, 2022, 42(24): 2428003.

[15] Zuo Y, Cao C F, Cao N P, et al. Optical neural network quantum state tomography[J]. Advanced Photonics, 2022, 4(2): 026004.

[16] Gao C K, Gaur P, Rubin S, et al. Thin liquid film as an optical nonlinear-nonlocal medium and memory element in integrated optofluidic reservoir computer[J]. Advanced Photonics, 2022, 4(4): 046005.

[17] Wang X D, Li R, Wang J, et al. One-dimension hierarchical local receptive fields based extreme learning machine for radar target HRRP recognition[J]. Neurocomputing, 2020, 418: 314-325.

[18] Hwang S W, Sugiyama J. Computer vision-based wood identification and its expansion and contribution potentials in wood science: a review[J]. Plant Methods, 2021, 17(1): 1-21.

[19] Toledo D, Córdoba-Jabonero C, Adame J, et al. Estimation of the atmospheric boundary layer height during different atmospheric conditions: a comparison on reliability of several methods applied to lidar measurements[J]. International Journal of Remote Sensing, 2017, 38: 3203-3218.

[20] Thomas R, Sylvain A, Tiago M. Deriving boundary layer height from aerosol lidar using machine learning: KABL and ADABL algorithms[J]. Atmospheric Measurement Techniques, 2021, 14(6): 4335-4353.

[21] Li H X, Chang J H, Liu Z X, et al. An improved method for automatic determination of the planetary boundary layer height based on lidar data[J]. Journal of Quantitative Spectroscopy and Radiative Transfer, 2020, 257: 107382.

[22] Liu Z X, Chang J H, Li H X, et al. Signal denoising method combined with variational mode decomposition, machine learning online optimization and the interval thresholding technique[J]. IEEE Access, 2020, 8: 223482-223494.

[23] Raghavendra K, Newsom R K, Berg L K, et al. On the estimation of boundary layer heights: a machine learning approach[J]. Atmospheric Measurement Techniques, 2021, 14(6): 4403-4424.

[24] Córdoba-Jabonero C, Sorribas M, Guerrero-Rascado J L, et al. Synergetic monitoring of Saharan dust plumes and potential impact on surface: a case study of dust transport from Canary Islands to Iberian Peninsula[J]. Atmospheric Chemistry and Physics, 2011, 11(224): 3067-3091.

[25] 刘诏. 基于CALIPSO星载激光雷达的边界层高度探测研究[D]. 北京: 中国科学院遥感与数字地球研究所, 2017.
Liu Z. Research on determination of PBLH based on CALIPSO space-borne lidar observations[D]. Beijing: Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, 2017.

[26] 何大义, 陈小玲, 许加强. 多属性群决策问题中基于最小叉熵的权重集成方法[J]. 控制与决策, 2017, 32(2): 378-384.
He D Y, Chen X L, Xu J Q. Weight aggregation method based on principle of minimum cross-entropy in multiple attribute group decision-making[J]. Control and Decision, 2017, 32(2): 378-384.

A Highly Robust Atmospheric Boundary Layer Height Estimation Method Combining K-means and Entropy Weight Method

Liu Zhenxing[1,2,3], Chang Jianhua[1,2*], Li Hongxu[4], Meng Yuanyuan[1], Zhou Mei[1], Dai Tengfei[1,2]

[1]*School of Electronics & Information Engineering, Nanjing University of Information Science & Technology, Nanjing* 210044, *Jiangsu, China*;

[2]*Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Nanjing University of Information Science & Technology, Nanjing* 210044, *Jiangsu, China*;

[3]*Department of Information Technology, Taizhou Polytechnic College, Taizhou* 225300, *Jiangsu, China*;

[4]*School of Electronic Information Engineering, Wuxi University, Wuxi* 214105, *Jiangsu, China*

Abstract

Objective The atmospheric boundary layer is the lowest layer of the troposphere, which is directly influenced by the surface. The atmospheric boundary layer height (ABLH) is an important parameter of the atmospheric boundary layer, whose value ranges from several hundred meters to thousands of meters. It plays an important role in analyzing the heat radiation transmission process in the boundary layer, acquiring the air pollution status, and formulating pollution control strategies. Lidar is an active remote sensing tool, which has high spatial and temporal resolutions and can continuously and automatically measure ABLH. The methods of estimating ABLH based on lidar data mainly include the threshold method, the gradient method, the wavelet covariance transform method, and the variance method. However, these methods are only suitable for specific meteorological conditions, and the interference of clouds or a suspended aerosol layer can easily lead to the misjudgment of ABLH. A highly robust ABLH estimation method combining K-means and entropy weight method, i. e., EK-means, is proposed to solve the problem of erroneous detection by commonly used lidar-based ABLH estimation methods under complex atmospheric structures. The proposed method improves the performance of ABLH estimation based on cluster analysis in terms of initial parameter selection and distance calculation. Compared with commonly used lidar-based ABLH estimation methods, the proposed method has a strong anti-interference ability. It can well track the diurnal variation process of the boundary layer under complex atmospheric structures. Under clear sky and cloudy weather or a suspended aerosol layer structure, the ABLH estimated by the proposed method is basically consistent with that measured by a radiosonde, and the correlation coefficient is 0.9718 and 0.9175, respectively. The proposed method has high robustness and can reliably estimate ABLH under different conditions.

Methods The proposed method integrates K-means and entropy weight method to improve the ABLH estimation performance based on cluster analysis from two aspects of initial parameter selection and distance calculation. Firstly, a sample dataset is constructed depending on the characteristics of the boundary layer, the free troposphere, a cloud layer, and a suspended aerosol layer. Then the utility function is introduced, and the entropy weight method is used to calculate the weight attributes of sample features. Next, the initial parameters of K-means are determined. The number n of intervals in the same direction is obtained by analyzing the gradient of the lidar backscattering signal, and the number of clustering categories ($k=n+1$ or $k=n+2$) can be obtained for different conditions. The initial center of clustering is selected as the position of the maximum signal intensity in the intervals in the same direction. Two centers are evenly selected in the first negative interval, and the Davis-Bouldin index is used for fine tuning. Finally, the ABLH is estimated with category features, which is located at the category boundary seeing the first decrease in the clustering strength from bottom to top.

Results and Discussions To assess the validity of the proposed EK-means, this paper uses the lidar data over Atmospheric Radiation Measurement (ARM) Southern Great Plains (SGP) central facility (C1) to estimate ABLH under various conditions. Experiments show the comparison results of the diurnal variation of ABLH tracked by four methods under the conditions of clear sky, polluted weather, and cloudy weather or a suspended aerosol layer structure (Figs. 5-7). The improved K-means and the proposed EK-means can reliably track the diurnal variation process of ABLH under these three conditions, and the proposed EK-means has the best performance (Figs. 5-7). The gradient method and the wavelet covariance transform method are susceptible to complex atmospheric structures such as clouds or a suspended aerosol layer, and the tops of clouds or the suspended aerosol layer is estimated as the ABLH, which has a large error (Fig. 7). Experimentally, the paper also compares the ABLHs estimated by the four lidar-based methods and by the radiosonde under clear sky and cloudy weather or a suspended aerosol layer structure (Figs. 8-9). The ABLH estimated by the proposed method under clear sky and cloudy weather or a suspended aerosol layer structure is consistent with that

measured by a radiosonde, and the correlation coefficients are 0.9718 and 0.9175, respectively [Fig. 8(d) and Fig. 9(d)]. The improved K-means also yields good experimental results with correlation coefficients of 0.9522 and 0.7986, respectively [Fig. 8(c) and Fig. 9(c)]. The ABLHs estimated by the gradient method and the wavelet covariance transform method are significantly different from that measured by a radiosonde under cloudy weather or a suspended aerosol layer structure, and the correlation coefficients are both less than 0.5 [Fig. 9(a) and Fig. 9(b)]. The proposed method has high robustness and can reliably estimate ABLH under different conditions (Table 1).

Conclusions The experimental results show that the proposed method is a highly robust ABLH estimation method compared with other commonly used lidar-based ones such as the gradient method and the wavelet covariance transform method. The proposed method can better track the diurnal variation of ABLH under clear sky, polluted weather, and cloudy weather or a suspended aerosol layer structure. Under the conditions of clear sky and cloudy weather or a suspended aerosol layer structure, the ABLH estimated by the proposed method has better consistency with that measured by a radiosonde, having a higher correlation coefficient and a smaller mean absolute error.

Key words remote sensing; lidar; atmospheric boundary layer height; complex atmospheric structures; cluster