基于轨迹聚类的天光光谱特征分析

蔡江辉[1]，杨雨晴[1]，杨海峰[1*]，罗阿理[2]，孔 啸[2]，张继福[1]

1. 太原科技大学计算机科学与技术学院，山西 太原 030024
2. 中国科学院国家天文台光学天文重点实验室，北京 100012

摘 要 天光背景扣除是LAMOST 1D光谱数据处理中重要的环节，其扣除好坏直接影响光谱产品质量，因此构造理想的超级天光光谱模型具有重要的意义。通常超级天光是由与目标天体同时观测的天光光纤光谱构造而成，同一区域的天光背景可能随着不同的观测时刻有着规律性的变化特征(如月相变化)，如果能充分分析并利用这些特征，可有效校正超级天光模型，从而提高减天光效果。轨迹聚类方法是一种分析目标随时、空变化特征的有效工具，针对LAMOST天光光谱中可能存在的变化规律，给出一种基于轨迹聚类的天光光谱特征分析方法。主要分以下三部分：首先是天光光谱的时序化描述。LAMOST pipeline采用且提供了每个观测天体的即时超级天光光谱，为了获取特定天区背景天光的光变特征，需选择天光光纤光谱以及扣除目标天体光谱的背景光谱，以5°视场(LAMOST望远镜视场)为单位，按观测日期MJD均匀分组，从而对特定区域的天光光谱进行了时序化表征；其次给出基于密度的天光光谱数据聚类算法STK-means。为解决随机参数导致收敛及聚类效果不理想的问题，在分析天光光谱时序数据特征的基础上，给出基于密度的相似性度量公式，并作为传统k-means聚类的初始参数选择依据，从而给出基于密度的天光光谱数据聚类算法STK-means；最后进行实验分析。实验验证了该方法的正确性和有效性以及不同初始参数K值的选择对聚类结果的影响。在此基础上，利用STK-means聚类方法，对LAMOST第一期巡天中一个完备小天区的天光光谱时序数据进行了轨迹特征分析，结果表明，除个别光谱质量较差或常说异常外，该特定区域的天光背景以农历每月十五、十六为中心向两边呈对称分布，反映了该区域观测过程中受月相的影响变化情况，该特征经量化后可为校正超级天光模型提供一种有效途径。同时，由于时序化描述过程中均匀采样的要求，该方法可适用于反银心、盘、晕等高天体数密度区域，而对于高银纬低数密度区域则需要更长时间的巡天观测。此外，该方法还可有效发现特定区域的离群(异常)天光光谱，为天文学家进一步分析提供珍稀样本。

关键词 天光背景；轨迹聚类；多目标光纤光谱；郭守敬望远镜(LAMOST)
中图分类号：P14 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)04-1301-06

引 言

2017年6月，LAMOST[1-2](large sky area multi-object fiber spectroscopic telescope)圆满完成了为期五年的第一期光谱巡天任务，共获取了约900万的光谱数据，为研究银河系及一般星系的形成与演化提供了有力的基础性数据。减天光是光谱数据预处理的重要环节，提高减天光精度有利于获得更高信噪比的光谱，从而对光谱进行深入分析。

为了扣除光谱中的天光成分，LAMOST通常利用一些光纤对天光进行采样，然后利用天光采样数据拟合出一个“超级天光”光谱近似作为特定区域的天光成分从目标光谱中扣除。这种方法忽略了有限的采样光纤与实际天光分布的差异，使得多目标光纤光谱的减天光精度不高。此外，也有利用B样条曲线拟合[3-4]，主成分分析[5-6]，滤波[7-8]，模板匹配[9-10]，等手段对减天光方法展开了研究。上述方法均在一定程度上提高了减天光精度，但其没有考虑天光流量的时序变化特征。

轨迹聚类[11-13]是时空数据处理的典型分支，其通过聚类的方法将行为相似的轨迹点聚为一簇，进而揭示轨迹点的时空分布规律和运动者的行为模式。为了研究LAMOST天光背景的变化规律，本文将轨迹聚类的思想引入天光光谱的处

收稿日期：2018-10-09，**修订日期**：2019-02-28
基金项目：国家自然科学基金项目(U1731126，61572343)，太原科技大学博士启动基金项目(20162007)资助
作者简介：蔡江辉，1978年生，太原科技大学计算机学院教授 e-mail：jianghui@tyust.edu.cn
*通讯联系人 e-mail：hfyang@tyust.edu.cn

理中。首先，选择天光光纤光谱以及扣除目标天体光谱的背景光谱，构造天光背景时序数据；其次，引入密度函数，改进 K-means 方法初始聚类中心的选择方式提出了天光背景聚类算法 STK-means；最后，利用 LAMOST 完备小天区范围内的天光背景时序数据，构造天光背景轨迹数据，并利用 STK-means 算法对天光背景轨迹数据进行聚类分析。

1 轨迹数据及其聚类

1.1 轨迹数据

轨迹数据[11]描述的是一个或多个移动对象运动过程的一种时空数据。一条移动对象的时空轨迹(如图 1 所示)可以形式化表述为

$$\mathrm{Tra}[id]=\{P_1,P_2,\cdots,P_i,\cdots,P_n\}$$

其中 n 为数据点数目，$P_i=\{(X_i,Y_i),T_i\}$，$1\leqslant i\leqslant n$，且 $T_{i+1}>T_i$。其语义表述为：移动对象在 T_i 时刻到达(X_i,Y_i)所示的位置。

如图 1 所示，移动对象的位置随时间发生变化，其中一些点相对集中，这些点可能为重要的地理位置。通过聚类可以将这些重要的点聚为一类，从而为后续分析移动对象的运动或行为模式提供依据。

图 1 移动对象的移动轨迹

Fig. 1 The trajectory of move object

1.2 天光背景时序数据

天光流量也会随着时间变化，将天光流量与轨迹中的位置对应，当天光流量为某个值时认为天光处于某种状态。本文将天光流量变化轨迹定义为

$$L\mathrm{Tra}[id]=\{LP_1,LP_2,\cdots,LP_i,\cdots,LP_n\}$$

其中，n 为天光流量轨迹数据点数目，$LP_i=\{(X_1,X_2,\cdots,X_m),T_i\}$，$1\leqslant i\leqslant n$，且 $T_{i+1}>T_i$，$(X_1,X_2,\cdots,X_m)$为 T_i 时刻的 m 维流量值序列，也即在 T_i 时刻天光处于$(X_1,X_2,\cdots,X_m)$所示的状态。天光流量变化轨迹是一种时序数据，因此，首先需要获得天光背景时序数据。

天光背景时序数据描述如下：

(1) 选择天光光纤光谱以及扣除目标天体光谱的背景光谱，读取其头文件信息，并以光纤为单位按光谱对应的 MJD 时间将光谱分组。

(2) 将所有按时间分组的天光光谱按时间顺序排列，并将光谱流量值归一化到同一尺度下得到天光背景时序数据集。

1.3 天光背景数据聚类

轨迹聚类主要目的是通过某种相似性度量尽可能将时空范围内相似的轨迹点划分到一个簇。本文采用欧式距离度量数据点的相似性，当数据点间的距离足够近时，认为其满足相似性条件可以划分到同一个簇。天光背景时序数据点的欧氏距离为

$$\mathrm{Dis}(LP_i,LP_j)=\sqrt{(x_{i1}-x_{j1})^2+\cdots+(x_{im}-x_{jm})^2} \tag{1}$$

K-means 算法是基于原型的经典聚类算法，该算法聚类过程简单快捷，适用于处理数据量大、维度高的数据集。但算法对初始聚类中心敏感，为了提升算法的聚类效率，本文利用密度函数度量数据点的集中程度并以此来确定初始聚类中心。数据点 LP_i 的密度计算方式如式(2)

$$\begin{cases}\rho_i=\sum\limits_{LP_j\in NLP_i(\mathrm{MeanDis})}\mathrm{e}^{-\left(\frac{\mathrm{dis}(LP_i,LP_j)}{\mathrm{MeanDis}}\right)^2}\\ \mathrm{MeanDis}=\dfrac{1}{C_n^2}\sum\limits_{i,j=1}^{n}\mathrm{dis}(LP_i,LP_j)\end{cases} \tag{2}$$

其中，$LP_j\in NLP_i(\mathrm{MeanDis})$表示点 LP_j 在 LP_i 的 MeanDis 邻域范围内。LP_i 的密度越大则其为中心点可能性越大。综上所述，本文天光背景数据聚类算法(STK-means)聚类过程如下：

INPUT：数据集 D，参数 K

OUTPUT：K 个簇 BEGIN：

1) 按式(1)计算点的距离 $\mathrm{Dis}(LP_i,LP_j)$；

2) 按式(2)计算各点的密度并将密度值降序排列；

3) 从排列后的密度值序列中分别选择 ρ_1，…，$\rho_{\lceil n/K\rceil(k-1)+1}$，…，$\rho_{n-\lceil n/K\rceil+1}$，$(1\leqslant k\leqslant K)$位置所对应的数据点作为 K 个聚类中心；

4) 将数据对象划分到其最近中心所在的簇；

5) 对于每个簇$(1\leqslant k\leqslant K)$重新计算簇中心；重复 4)和 5)直到每个簇的中心不再变化。

END

2 实验部分

本文实验平台配置为 Intel core i5 3470 CPU，4G 内存，64 位 Win 7 OS，程序实现工具 JAVA。

2.1 完备小天区天光背景轨迹数据及其聚类

LAMOST 天光光谱覆盖范围广，针对 LAMOST 所有天光背景时序数据聚类的时空开销很大，不利于发现天光背景时序变化特征。本文以天光背景时序数据集为基础，将特定天区范围内的所有天光时序数据抽取出来，构成该天区范围内的天光背景轨迹数据，然后利用 STK-means 对其进行聚类。

LAMOST 指向区域的完备光谱观测(the LAMOST complete spectroscopic survey of pointing area，LCSSPAR)项目[14]旨在完成两个天区内的所有河内和河外源的光谱观测。这两个天区的中心坐标分别为 RA＝37.881 509 39°，DEC＝3.439 345 00°(天区 A)和 RA＝21.525 988 792°，DEC＝－2.200 949 833°(天区 B)。本文以完备小天区 A 为依据研究其 5°视场范围内天光背景随时间的变化情况。

天区 A 的天光背景轨迹数据构造过程如下：

(1) 以天光时序数据为基础，将其所有天光光谱的赤经赤纬与天区 A 的中心交叉，提取出天区 A 的 5°视场范围内所有天光时序数据的分组。本文共获得了 2011 年 10 月至 2016 年 11 月的 803 条天光 Fit 文件，从天光背景时序数据中共提取出了 159 个分别编号为 1—159 的不同时间点的分组，其中一共包含了以光纤为单位的209 752条天光光谱。

(2) 从每个时间点分组中选择 3 条(不同时间点分组的光谱条数存在较大差异，最大分组包含 5 378 条光谱，最小的分组包含 3 条光谱。本文以仅包含 3 条光谱的时间点为基础，将其中的 3 条光谱作为基础光谱，其余各时间点的 3 条光谱分别位于基础光谱 3 角秒范围内)光谱构成该时间点下的天光流量面。如图 2 所示，图 2(a)，(b)和(c)分别为 2013 年 10 月 14 日的一个时间点分组中 3 条光谱，图 2(d)为(a)，(b)，(c)三条光谱拼接成的该时间点的流量面。

(3) 以天光流量面为一个轨迹数据点，所有时间点下的天光流量面按时间顺序排列得到天区 A 对应的一条天光背景轨迹数据(图 3)。

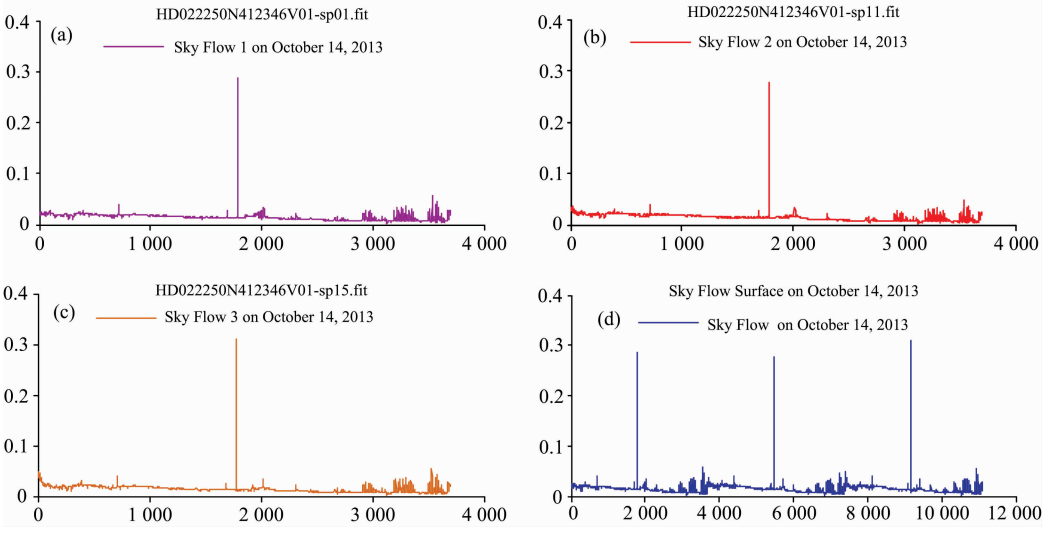


图 2　2013 年 10 月 14 日的天光流量面

Fig. 2　Skylight flow surface on October 14, 2013

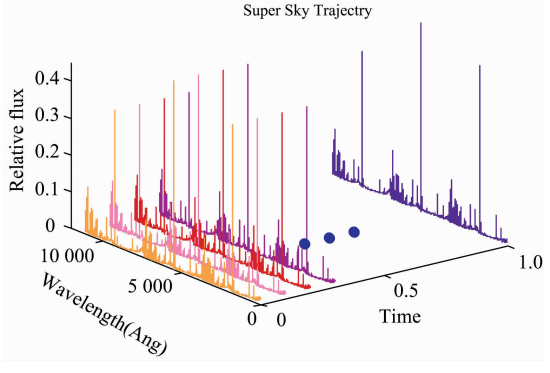


图 3　天光背景轨迹数据

Fig. 3　Skylight background trajectory data

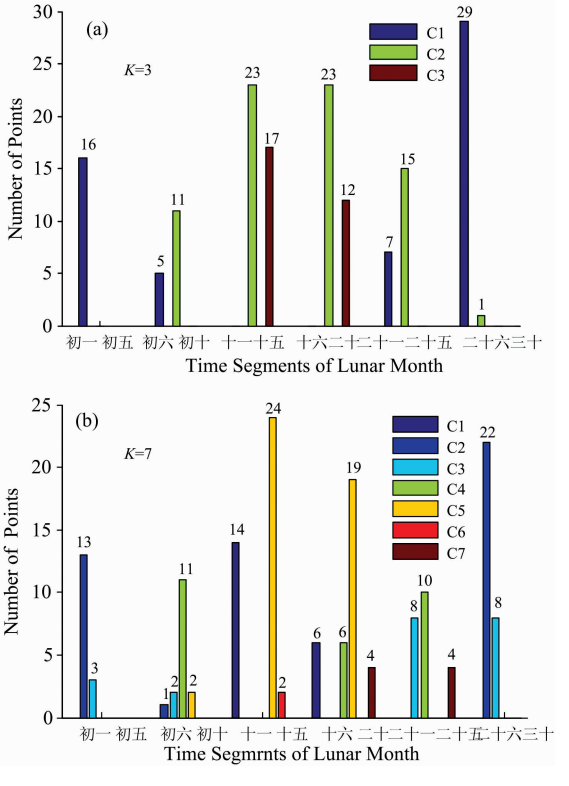
如图 3 所示，每条不同颜色标记的天光流量曲线是由某个时间点下的三条天光流量拼接形成的天光流量面。以流量面为处理单元，利用 STK-means 算法聚类天光背景轨迹数据并对其结果进行分析。

3　结果与讨论

改进后的 STK-means 算法涉及到的参数为 K，本文在多组 K 值下对天区 A 的天光背景轨迹数据进行了聚类分析，实验结果及相关分析如下。

3.1　天光流量面的对称分布趋势

图 4(a)，(b)，(c)为 K 值取 3，7，9 时天光光谱对应农历时间的分布情况。图 4(a)中 C1，C2 和 C3 为簇标号，分别用蓝、绿、褐表示。初一到初五这段时间的流量面全部分布在 C1，二十六到三十这段时间，除 1 个流量面分到 C2 外几乎所有流量面都被分到 C1。初六到初十中有 5 个流量面被分到了C1，11个流量面被分到了C2，二十一到二十五中7



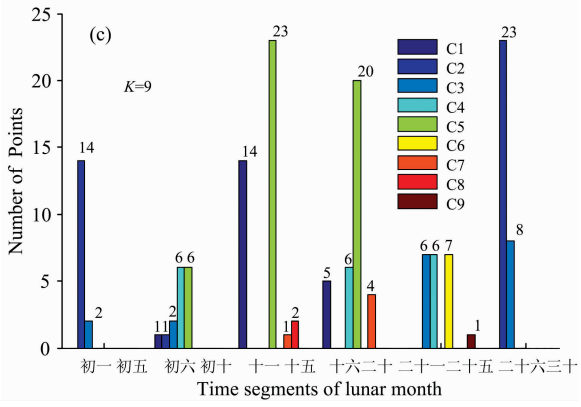


图 4 (a)，(b)，(c)分别为 K 值取 3，7，9 的聚类结果

Fig. 4 (a)，(b)，(c) are clustering results when K takes 3，7，9

个流量面被分到了 C1，15 个流量面被分到了 C2。十一到十五、十六到二十这两段时间的流量面均只被分到了 C2 和 C3 中。上述流量面以十五、十六为中心呈现出明显对称分布规律，为进一步验证该分布规律，本文调整 K 值进行了实验。

图 4(b)为 $K=7$ 时流量面的分布情况，图中仅展示了包含的流量面数目不为 0 的簇。由于篇幅限制，仅给出了 K 为 7 的统计结果，$K=4$，5，6 的分布结果与 $K=3$ 时类似。图 4(b)中初一到初五的流量面只分布在 C2 和 C3 中，其余 5 个簇中不包含初一到初五这个时间段的流量面。二十六到三十这段时间的流量面只分布在 C2 和 C3 中。其余时间段的对称分布趋势减弱，对称分布趋势减弱主要是由于簇数目增加使得边缘点数目增多造成的。总体来说，初六到初十、二十一到二十五的流量面大部分集中在 C4 中，十一到十五、十六到二十的流量面大部分集中在 C5 中。从上述分析可得，天光流量基本以农历每月十五、十六为中心呈现出左右对称分布趋势，与月相分布规律相近。

3.2 异常流量面分析

图 4(b)中的簇 C6 仅包含 2 个流量面(编号为 19 和 71，其 MJD 时间分别为80 971 004，81 564 379)，且 C6 在所有时间段中仅出现 1 次。为深入分析 19 和 71 号流量面，将 K 值设置为 8 和 9 进行实验。图 4(c)为 $K=9$ 时的聚类结果。图 4(c)中 C8 包含了 19 和 71 号流量面，在图 4(c)中这两个流量面也单独成一簇且仅分布在十一到十五这个时间段中。

图 5 为 17 和 19 号流量面的光谱。图中这两个流量面的光谱很相似，易被划分到同一个簇，而这两个天光流量面与实际天光存在较大差异，因此，它们单独成一簇。

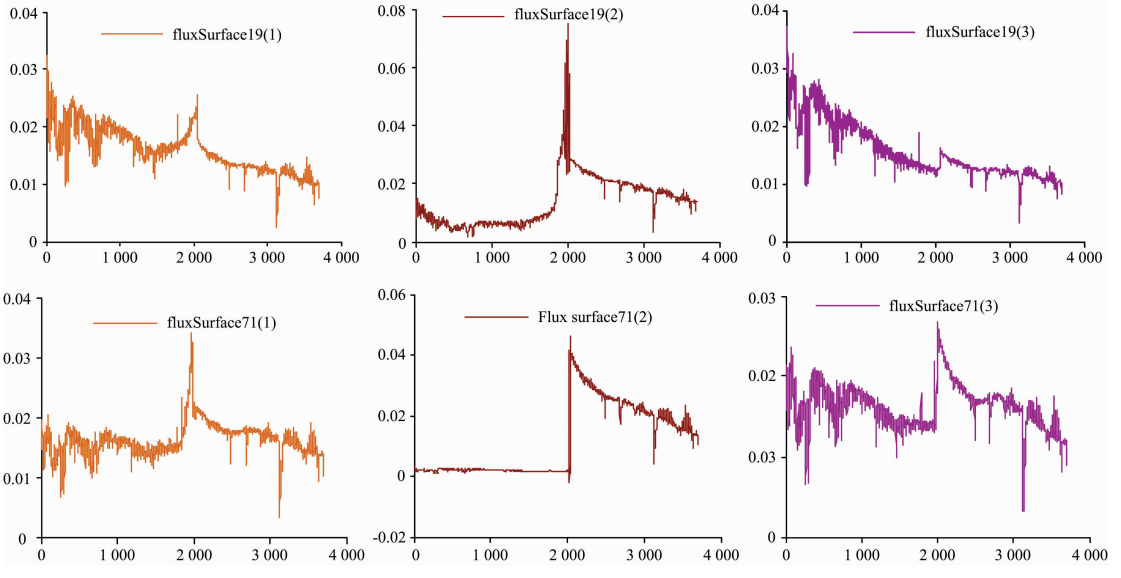


图 5 编号为 19 和 71 的流量面光谱

Fig. 5 Flow surface spectra numbered 19 and 71

图 4(c)中簇 C9 也仅包含一个流量面，且仅出现在时间段二十一到二十五中。特殊流量面可能是某些原因导致的坏点，且随着簇数目的增多可能发现更多特殊流量面。根据上述猜想本文增加 K 值进行实验。实验发现，随着 K 增加(表 1)发现了更多异常流量面。当 K 为 30 时发现了编号为 2，3，16，19，28，33，71，73，90，110，122，137，143 和 144 的共 14 个异常流量面，部分异常流量面如图 6 所示。图 6 中编号为 3 和 90 的流量面对应分组的 MJD 时间为80 465 775，81 959 279。上述光谱图像与正常天光光谱存在较大差异，它们可能产生于人为计算错误、设备仪器的故障等，聚类中这些特殊光谱与正常光谱存在差异较大被视为坏点单独成簇。

表 1 K 取不同值时发现的异常流量面数量

Table 1 Number of abnormal flow surfaces found at different K values

K	异常流量面数	K	异常流量面数	K	异常流量面数
3	0	11	3	19	4
5	1	13	3	21	6
7	1	15	4	25	10
9	2	17	5	30	14

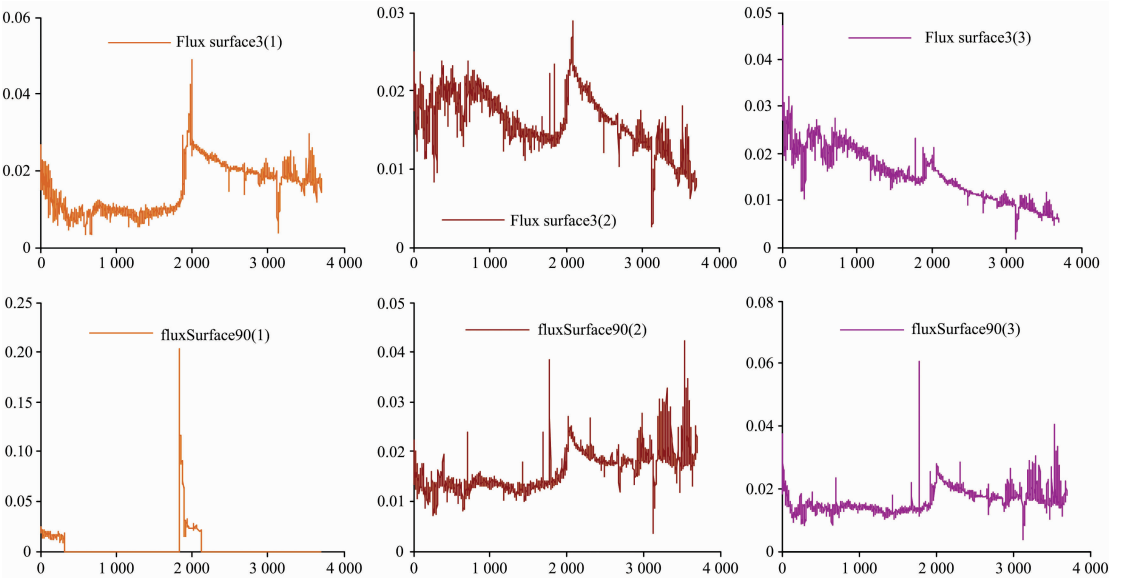


图 6　编号为 3 和 90 的流量面光谱

Fig. 6　Flow surface spectra numbered 3 and 90

4　结　论

针对天光的时序变化特征，将轨迹聚类的有关思想引入到天光光谱的分析中，从时序数据分析的角度出发对天光变化规律进行了分析。实验结果表明，天光背景大致以农历每月十五、十六为中心呈对称分布趋势，这种分布特征经量化后为校正超级天光模型提供了一种新的途径。同时，由于时序化描述过程中均匀采样的要求，该方法可适用于反银心、盘、晕等高天体数密度区域，而对于高银纬低数密度区域则需要更长时间的巡天观测。此外，该方法还可有效发现特定区域的离群(异常)天光光谱，为天文学家进一步分析提供珍稀样本。

References

[1] Luo A L, Zhao Y H, Zhao G, et al. Research in Astronomy and Astrophysics, 2015, 15(8): 1095.
[2] Cui X Q, Zhao Y H, Chu Y Q, et al. Research in Astronomy and Astrophysics, 2012, 12(9): 1197.
[3] Fischer J L, Bernardi M, Meert A. Monthly Notices of the Royal Astronomical Society, 2017, 467(1): 490.
[4] Blanton M R, Kazin E, Muna D, et al. The Astronomical Journal, 2011, 142(1): 31.
[5] Bai Z R, Zhang H T, Yuan H L, et al. Research in Astronomy and Astrophysics, 2017, 17(9): 91.
[6] Bai Z, Zhang H, Yuan H, et al. Publications of the Astronomical Society of the Pacific, 2017, 129(972): 024004.
[7] Griesbach J, Wetterer C, Sydney P, et al. Efficient Photometry in-Frame Calibration (EPIC) Gaussian Corrections for Automated Background Normalization of Rate-Tracked Satellite Imagery. Advanced Maui Optical and Space Surveillance Technologies Conference. 2015.
[8] Bland-Hawthorn J, Englund M, Edvell G. Optics Express, 2004, 12(24): 5902.
[9] Fernandes M V, Horns D, Kosack K, et al. Astronomy & Astrophysics, 2014, 568: A117.
[10] AN Ran, PAN Jing-chang, YI Zhen-ping, et al(安　冉，潘景昌，衣振萍，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2017, 37(1): 273.
[11] Yuan G, Sun P, Zhao J, et al. In: Proc. of the Artificial Intelligence Review, 2017, 47: 123.
[12] Luo T, Zheng X W, Xu G L, et al. International Journal of Geo-Information, 2017, 6(3): 63.
[13] Guan B, Liang X L, Chen J Y. TELKOMNIKA Indonesian Journal of Electrical Engineering, 2013, 1(1): 115.
[14] Wu H, Yang M, Lam M I, et al. Proceedings of Symposium of the International-Astronomical-Union, 2016, 11(S317): 371.

Spectral Analysis of Sky Light Based on Trajectory Clustering

CAI Jiang-hui[1], YANG Yu-qing[1], YANG Hai-feng[1*], LUO A-li[2], KONG Xiao[2], ZHANG Ji-fu[1]

1. School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China
2. Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China

Abstract Skylight background subtraction is an important part of LAMOST 1D spectral data processing, and constructing ideal super sky spectral models is of great significance since it may directly affect the quality of the spectral products. Generally, the super sky spectral models are composed of the spectra from sky fibres simultaneously observed with target objects, and sky background may be of regular variation along with different observation times. Taking full account of these timing features, the super skylight model can be effectively corrected to improve the skylight reduction effect. Meanwhile, the trajectory clustering method is an effective tool for analyzing the characteristics of the target with temporal and spatial variation. Therefore, a method for analyzing the characteristics of the sky spectra based on the trajectory clustering is provided in this paper orienting to the possible variation laws in the sky spectra of LAMOST. It includes the following 3 parts: (1) the time series description of sky spectra. In fact, LAMOST pipeline uses and provides the instant super sky spectra for each observed target. In order to obtain the light-changing characteristics of the sky background spectra of a specific sky area, the time series of sky spectra are re-described by selecting the sky fiber spectra and background spectra without target component, taking the 5-degree field of view (the Fov of LAMOST) as processing unit, and evenly grouping these spectra by observation date. (2) density-based clustering algorithm (STK-means) for sky spectra. In order to solve the problem that the random parameters may lead to relatively poor convergence and clustering, a density-based similarity measurement formula is studied. The values of this formula are used as the selection basis of the initial parameters, and then a new algorithm named STK-means is proposed after updating the traditional *k*-means algorithm. (3) experiment analysis. Firstly, by experiment, the correctness and effectiveness of this method is verified, and clustering effect is analyzed by utilizing different initial parameter k. And then, the trajectory characteristics of sky spectral time series are analyzed by selecting the sky spectra from one of complete small sky areas in the first phase of LAMOST survey. The experimental results show that the sky background in particular region is distributed symmetrically around the lunar 15th and 16th of each month, which indicates the influence partly from the moon phase during the observation process in this sky area. These timing characteristics can be quantified to correct the super sky spectral model. Meanwhile, uniform sampling of data during the description of time-series spectra is very important, so this method can be effectively applied to the regions of high celestial number density such as GAC, disk, halo, etc. On the contrary, the longer time survey is necessary for the low number density areas. In addition, this method may also effectively find outlier sky spectra of specific regions, which will provide rare samples for further physical study.

Keywords Sky background; Trajectory clustering; Multi-object fibre spectroscopy; LAMOST

(Received Oct. 9, 2018; accepted Feb. 28, 2019)

* Corresponding author