Research in Astronomy and Astrophysics, 25:057001 (7pp), 2025 May

https://doi.org/10.1088/1674-4527/adcb8b
CSTR: 32081.14.RAA.adcb8b

# LAMOST Spectral Data Processing: Classification, Redshift Measurement, and Data Product Creation

Xiao Kong[1] and A-Li Luo[1,2]
[1] Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, China; kongx@bao.ac.cn
[2] University of Chinese Academy of Sciences, Beijing 100049, China
*Received 2024 November 15; revised 2025 February 23; accepted 2025 February 26; published 2025 May 20*

## Abstract

The Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST) has become a crucial resource in astronomical research, offering a vast amount of spectral data for stars, galaxies, and quasars. This paper presents the data processing methods used by LAMOST, focusing on the classification and redshift measurement of large spectral data sets through template matching, as well as the creation of data products. Additionally, this paper details the construction of the Multiple Epoch Catalogs by integrating LAMOST spectral data with photometric data from Gaia and Pan-STARRS, and explains the creation of both low- and medium-resolution data products.

*Key words:* catalogs – methods: data analysis – techniques: spectroscopic – surveys

## 1. Introduction

The Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST; Xiang-Qun et al. 2012) has become a cornerstone in astronomical surveys, providing a wealth of spectroscopic data to the scientific community. Over the course of its operation, LAMOST has released ten sets of data products (Li et al. 2012), covering millions of spectra for stars, galaxies, and quasars. These data products have been instrumental in a wide range of astrophysical research, from stellar population studies to extragalactic astronomy.

LAMOST utilizes template matching for spectral classification and redshift measurement, initially drawing from the SDSS spectral templates (Adam et al. 2012). After the first data release (DR1), we constructed 183 stellar spectral templates using high-quality spectra obtained by LAMOST (Peng et al. 2014). Later, due to the low accuracy of luminosity classification for A-type stars, we removed luminosity classifications and merged duplicate A-type stellar templates. Based on the research by Jing et al. (2019), the M-type stellar templates were updated to include 20 templates with labels for giants and dwarfs. For each data release, similar to the SDSS spectral data release policy, we adhere to the principle that each new data release (DR) includes data from previous releases.[3] However, if there are discrepancies caused by updates in the processing pipeline, the new DR data may not be entirely consistent with the older versions. In such cases, detailed explanations are provided on the release website.

This paper introduces the methods used in LAMOST's data processing pipeline, focusing on the classification and redshift measurement of large spectroscopic data sets through template matching, as well as the creation of data products.

[3] https://www.lamost.org/lmusers/cms/article/view?id=1

## 2. Spectral Classification and Redshift Measurement

The LAMOST pipeline draws inspiration from the SDSS pipeline's spectral classification method (Hutchinson et al. 2016), using template matching to classify the vast amount of spectroscopic data and measure redshifts. In this paper, the "redshift" refers to the parameter $z$ in the Pipeline program. If we are considering the radial velocity of a star in a physical sense, the value of $z$ needs to be calculated using the speed of light. In all subsequent content, we will no longer distinguish between radial velocity and redshift, and will only consider the parameter $z$ in terms of numerical calculations, uniformly referring to it as redshift.

LAMOST currently has three major categories of spectral templates: stars, galaxies, and QSOs. Among these, the stellar templates are the most extensive, having undergone several iterations of updates. There are now 165 stellar templates, including main-sequence stars and special types of stars.

### 2.1. Template Fitting

Template matching is the primary method employed by LAMOST to classify celestial spectroscopic and determine their redshifts. The approach uses a set of well-defined templates representing various types of celestial objects, including stars, galaxies, and quasars. These templates cover a wide range of spectral types, including stars (O, B, A, F, G, K, M, white dwarfs, carbons, and their subclasses), galaxies, and quasars. In this context, classes refer to the three broad categories of stars, galaxies, and quasars, while subclasses,

such as A1, F3, or G8, further refine the classification. The letter represents the stellar type, and the numbers from 0 to 9 denote finer subdivisions within that type.

The classification process involves using templates to classify spectra captured by the telescope and simultaneously determine the redshift of each source. Since observed spectra often have inaccuracies in flux calibration, the template continuum must be adjusted before matching, ensuring the continuum shapes of the observed and template spectra align. This adjustment allows line features to have maximum weight in the $\chi^2$ distance calculation, resulting in more accurate matching.

The continuum adjustment process uses singular value decomposition (SVD) to derive characteristic vectors, which are then used to modify the template continuum. Specifically, let $A = (f, x_0, x_1, x_2, x_3)$, where $f$ is the template spectrum, and $(x_0, x_1, x_2, x_3)$ are polynomials of degrees 0, 1, 2, and 3, respectively, each having the same length as the template spectrum. The matrix $AA^T$ undergoes SVD, resulting in $AA^T = U\Sigma V^T$, where $A^T$ represents the transpose of $A$. The coefficient vector $c$ is then calculated as $c = \frac{V}{\Sigma}U^T(A \times s)$, where $s$ is the observed spectrum vector. Finally, $A \times c$ gives the modified template spectrum with an adjusted continuum.

For stellar templates, the redshift search range is typically very narrow, from $-0.004$ to $+0.004$, with a step size of 0.0002. This means that the $\chi^2$ distance between the observed spectrum and the template is computed at redshift $-0.004$, and the template is shifted by 0.0002 at each step until it reaches $+0.004$. This process yields a $\chi^2$ vector for each template. For galaxies, the redshift range is from $-0.0017$ to $+1$, with a step size of 0.0007, while for quasars, the redshift range is from $-0.0033$ to $+7$, with a step size of 0.003. Once the $\chi^2$ matrices for all templates are computed, the best match is found by selecting the template and redshift corresponding to the minimum $\chi^2$ value.

After obtaining the best match, a confidence estimation program is used to evaluate the reliability of the template-matching result. If the confidence level is below 80%, it indicates that the continuum shape may not have been accurately adjusted, and a polynomial fitting method is employed for re-matching.

To further refine the matching results, several criteria are used to adjust the final selection based on the minimum $\chi^2$ value:

1. The final adjusted template will have five polynomial coefficients. After forming the $\chi^2$ matrix, the results are sorted in an ascending order of $\chi^2$ values. If the leading coefficient of the polynomial is negative, it indicates that the template spectrum has been inverted after continuum adjustment. If the other coefficients are all positive and the spectral type differences among the top four matched templates do not exceed ten subclasses, the best match is moved to the end of the list.
2. If the best-matching template is not a stellar template, but at least seven of the next nine best matches are stellar templates, it is assumed that the observed spectrum is stellar, and the non-stellar template is moved to the end of the list.
3. For the sorted $\chi^2$ sequence, the differences between adjacent matches are calculated. If the differences between several consecutive matches are less than 1% of the $\chi^2$ value, and their spectral types differ by no more than five subclasses, these matches are considered unreliable and are moved to the end of the list.

This detailed template-matching process, with continuum adjustment and iterative refinement, ensures highly reliable spectroscopic classification and redshift estimation for the vast LAMOST data set.

**Table 1**
Ranges of Weight $C$ for Different Types of Objects

| | $C$ |
|---|---|
| O | 0.90 |
| B | 0.90 |
| A | 0.74 |
| F | 0.58 |
| G | 0.42 |
| K | 0.26 |
| M | 0.10 |
| others | 0.50 |

### 2.2. Confidence of Fitting

Assessing the confidence of spectroscopic template matching is crucial for ensuring reliable classification and parameter estimation. Traditional methods relying solely on the minimum $\chi^2$ value do not always provide an accurate measure of the reliability of a match, particularly for spectra with varying signal-to-noise ratios or systematic flux calibration errors. Here, we introduce an empirical confidence estimation method that leverages the relative flux distribution at the blue and red ends of the spectrum.

The underlying rationale is that different stellar types exhibit distinct flux distributions in the blue and red wavelength regions. For example, O-type stars have significantly higher average flux in the blue end compared to the red end, whereas M-type stars show the opposite trend. To quantify this effect, we define a confidence level $\mathrm{Conf} = Cf(B) + (1 - C)f(R)$, where $f(B)$ and $f(R)$ are Gaussian-fitted residual distributions at the blue and red ends, and $C$ is a weight factor determined based on the spectral type (Table 1).

The blue and red fitting functions are calculated as follows:

$$f(B) = \frac{15 - \mu_B}{13}, f(R) = \frac{15 - \mu_R}{10}.$$

These formulas were derived empirically by fitting a large set of LAMOST spectra across different spectral types. Specifically, $\mu_B$ and $\mu_R$ represent the mean residuals at the blue and red ends, respectively, obtained from Gaussian fits. The coefficients in the equations were optimized through statistical analysis of a representative sample of hundreds of thousands of spectra, ensuring their effectiveness in characterizing the confidence of template matching within the LAMOST data framework. This empirical formulation accounts for systematic deviations in flux calibration and noise characteristics unique to LAMOST observations. The parameters (15, 13, and 10) were determined by iteratively adjusting the formula and validating the confidence estimates against manually classified spectra. The resulting confidence estimation scheme has been extensively tested and fine-tuned to optimize classification accuracy in LAMOST 1D Pipeline.

The weight factor $C$ is derived from the ratio of the mean flux in the blue and red regions for each spectral type. For example, in O-type stars, the blue-end flux is approximately 9 times that of the red-end flux, leading to $C = 0.9$. Similarly, for F-type stars, this ratio is about 1.5, and considering the relative number of spectral lines in the blue and red regions (approximately 3:2), we obtain $C \approx 0.6$. These values were initially estimated from statistical flux distributions and further refined through iterative testing.

To optimize $C$, we performed extensive validation using a sample of hundreds of thousands of LAMOST spectra across various spectral types, where the ground truth classifications were verified through manual inspection. The pipeline was iteratively run, and the statistical differences between the automated classification results and the true classifications were analyzed to fine-tune the weight values. The final set of weights represents an empirically optimized confidence estimation scheme tailored specifically for LAMOST data.

While this approach is non-standard, it has demonstrated significant effectiveness in distinguishing reliable and unreliable matches across a wide range of stellar spectra. The methodology is data-driven and designed to enhance classification accuracy by incorporating information beyond simple $\chi^2$ minimization.

### 2.3. Classification Performance

Spectroscopic classification can be significantly affected by the quality of the observed spectra. In particular, spectra with poor merging between the blue and red arms of the spectrograph often lead to misclassification in traditional template-matching algorithms. If one end of the spectrum is of significantly lower quality, it can bias the overall classification result, making it difficult to obtain reliable stellar type identification.

To evaluate the robustness of our pipeline against such challenges, we present two examples from LAMOST DR10, where poor merging at the blue and red ends might have hindered accurate classification. Figure 1 shows the observed spectra (black curves) and their corresponding best-matching templates (green curves). The obsids of these two spectra are 27501001 and 27501023, respectively. Despite their imperfections, the pipeline successfully classifies them as G3- and M1-type stars, respectively.

The success of our classification pipeline in these cases can be attributed to the confidence estimation approach described in Section 2.2. By incorporating residual-based confidence measures and balancing the contribution of different spectral regions using a weighting factor, our method reduces the impact of flux discrepancies at different wavelengths. In contrast, traditional $\chi^2$-based template-matching approaches often fail under these conditions due to the significant weight given to the poorly merged spectral region.

These examples demonstrate that the LAMOST spectral classification pipeline is robust against suboptimal spectral data and remains reliable under real-world observational conditions. The confidence estimation method plays a crucial role in ensuring the accuracy of classification results even when spectral merging is not ideal.

To further demonstrate the effectiveness of our classification pipeline, we present examples of high-quality spectra across a diverse range of astrophysical objects in Figure 2. This set includes spectra of various stellar types (O, B, A, F, G, K, M, white dwarfs, carbon stars and cataclysmic variables), as well as extragalactic sources such as galaxies and QSOs. Each spectrum is plotted in black, while the best-matching template is overlaid in light green.

Since each object has a different redshift, particularly for extragalactic sources, we have chosen not to correct for redshift in this figure to maintain visual alignment and clarity. Instead, the goal is to illustrate the effectiveness of the template-matching process across a wide range of objects. All spectra are plotted using relative flux, and they are arranged from top to bottom according to their spectral type.

These examples further validate the robustness of our pipeline in accurately classifying both stellar and extragalactic sources, demonstrating its reliability for LAMOST's spectral data processing. The consistency between the observed spectra and the best-matching templates across different object types highlights the effectiveness of the classification approach, even for objects with significant redshift. This figure, combined with the previously shown low-quality spectra examples, provides a comprehensive overview of the pipeline's performance under varying observational conditions.

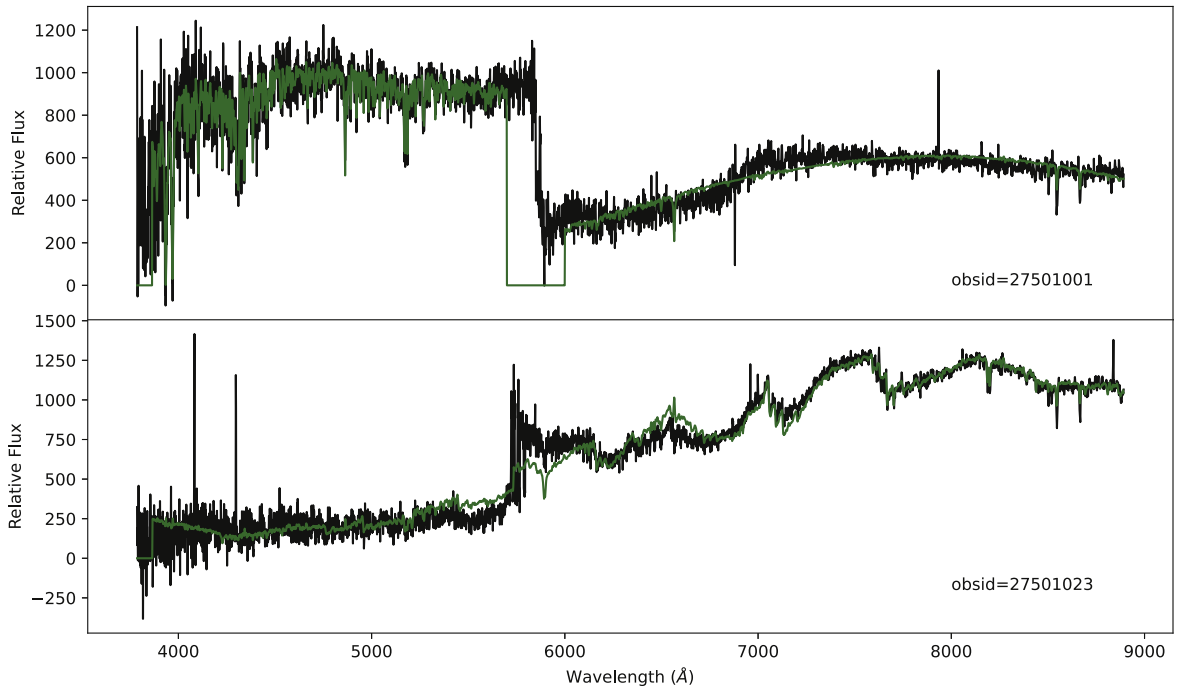


**Figure 1.** Example of spectroscopic classification for two low-quality spectra from LAMOST DR10. The black curves represent the observed spectra, while the green curves represent the best-matching templates. The pipeline successfully classifies the spectra despite poor merging at the blue and red ends.

## 3. Multiple Epoch Catalog

The construction of the homogeneous catalog, or Multiple Epoch Catalog, involves combining LAMOST spectroscopic data with photometric data from other major surveys, such as Gaia (Gaia Collaboration et al. 2016, 2023) and Pan-STARRS (Tonry et al. 2012; Heather 2018). Since LAMOST lacks its own photometric data, the integration with Gaia and Pan-STARRS is essential for providing a complete data set for astrophysical analysis. The cross-matching radius is determined by the observation accuracy of the LAMOST telescope fibers, and is set to $3''$ (Chen et al. 2014). All data are stored in a database for subsequent retrieval and analysis. The database tables contain four basic fields for each entry: TARGETID, OBJRA, OBJDEC, and SOURCE, which indicate the data ID, R.A., decl., and data origin, respectively.

The process of constructing the Multiple Epoch Catalog involves the following steps (Figure 3):

1. *Calculating Cross-Matching Within LAMOST Data*. The first step is to determine the internal cross-matching within the LAMOST data set. The spherical distance between any two spectra is calculated using the formula:

$$\arccos(\sin(\mathrm{dec1}) * \sin(\mathrm{dec2}) + \cos(\mathrm{dec1}) * \cos(\mathrm{dec2}) * \cos(\mathrm{ra1} - \mathrm{ra2}))$$

   where (ra1, dec1) and (ra2, dec2) are the coordinates of the two spectra. If the distance is no more than $3''$, the two spectra are considered to be from the same source.

2. *Obtaining Photometric Counterparts from Catalogs*. The next step is to obtain potential photometric counterparts for LAMOST data. Cross-matching with Pan-STARRS photometric data is performed using the coordinate cross-matching function available on the Pan-STARRS website. All photometric data matching LAMOST spectra within $3''$ are added to the database. Additionally, cross-matching with Gaia photometric data is performed using the TOPCAT tool, and all matching Gaia candidates are also stored.
3. *Constructing the Multiple Epoch Catalog*. For each celestial target observed by LAMOST, if there is exactly one matching Gaia photometric source, that Gaia source is considered the photometric counterpart of the LAMOST target, and a unique identifier (UID) is assigned as "g" plus the Gaia "source_id." If no Gaia source matches, but there is exactly one matching Pan-STARRS photometric source, that Pan-STARRS source is considered the photometric counterpart, and the UID is assigned as "p" plus the Pan-STARRS "objid." If multiple Gaia or Pan-STARRS sources match, or if no photometric source matches, the UID is assigned as "l" plus the LAMOST "obsid" from the first observation of the target.
4. *Constructing LAMOST Photometric Magnitudes*. Since the magnitude ranges of LAMOST low-resolution spectra align well with Pan-STARRS data, and medium-

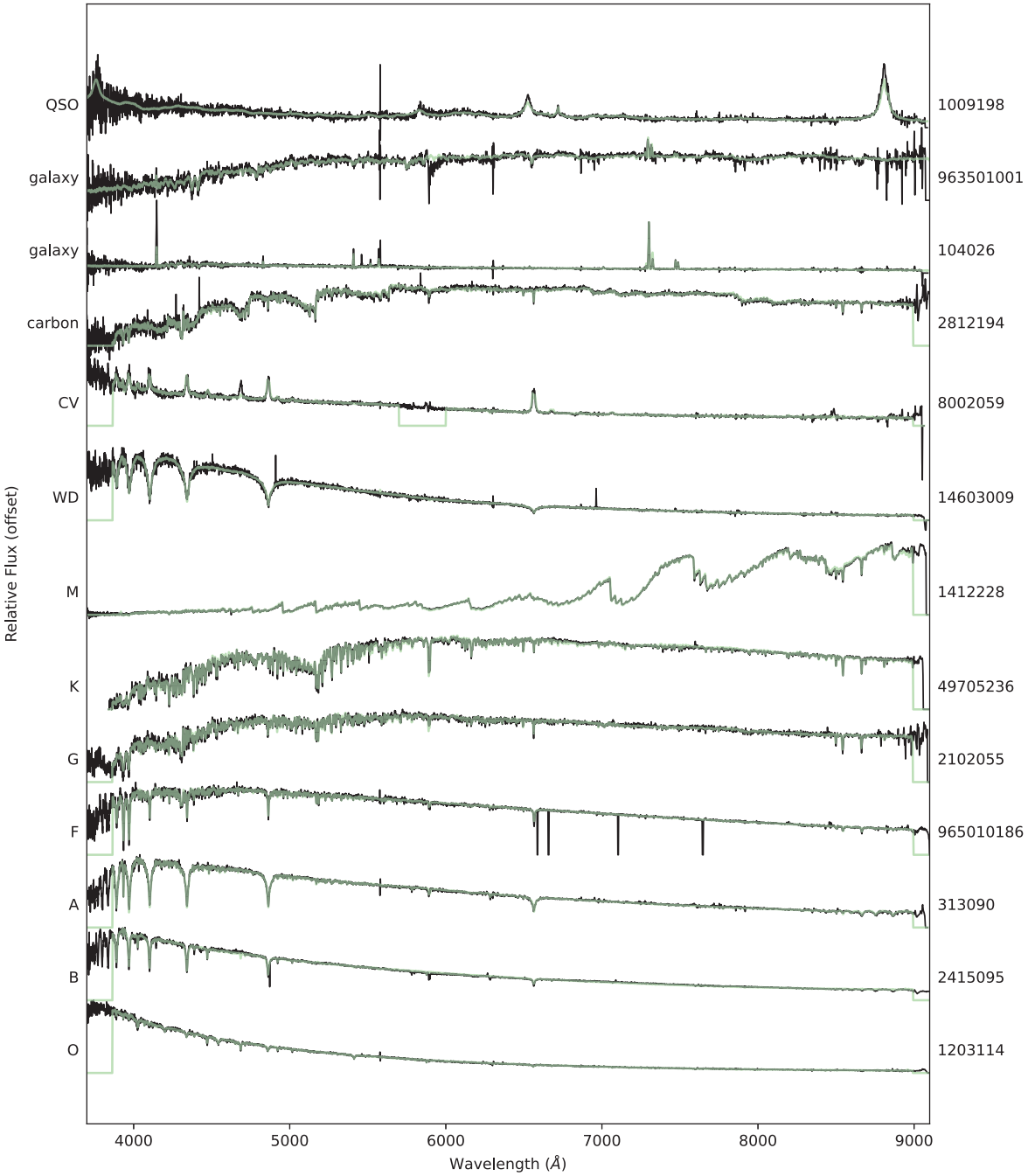


**Figure 2.** Example of spectral classification for a diverse set of astrophysical objects, including various stellar types, galaxies, and QSOs. The black curves represent the observed spectra, while the light green curves correspond to the best-matching templates. Spectral types are labeled on the left, and obsids are listed on the right.

resolution spectra (MRS) align well with Gaia data, the corresponding photometric magnitudes are used to assign photometric magnitudes for LAMOST data.

(a) The *grizy* magnitudes from the nearest Pan-STARRS photometric source within 3″ are used as the photometric magnitudes for LAMOST LRS.

(b) The BP and RP magnitudes from the nearest Gaia photometric source within 3″ are used as the blue and red photometric magnitudes for LAMOST MRS.

This method effectively integrates LAMOST, Gaia, and Pan-STARRS data, enhancing the precision and reliability of spectroscopic analyses, especially for those studies that require consistent photometric calibration.

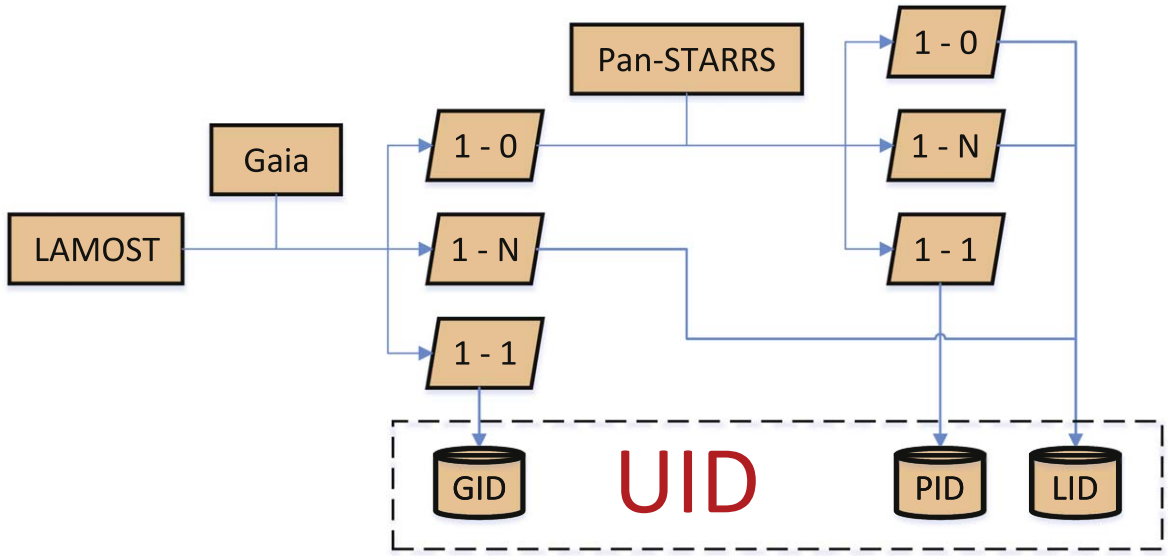


**Figure 3.** Process of constructing the Multiple Epoch Catalog. LAMOST spectra are first cross-matched with Gaia data. If there is exactly one corresponding Gaia photometric source, the source_id is used as the GID. If there is no matching Gaia source, the spectra are then cross-matched with Pan-STARRS. If there is exactly one photometric source, the objid from Pan-STARRS is used as the PID. In all other cases, the obsid from the first LAMOST observation is used as the LID. The GID, PID, and LID are standardized into a unique identifier (UID), which serves as the sole identifier in the multiple epoch catalog.

## 4. Released Data

LAMOST provides a variety of data products, catering to different observational and analytical needs. The data products are divided into LRS and MRS categories, each with its unique features and data formats.[4]

### *4.1. Low-resolution Spectroscopic Data Products*

The LRS data products include spectroscopic FITS files, spectroscopic images in PNG format, catalog files, and sky files. The FITS files contain detailed information about each observation, divided into several sections:

1. *FILE INFORMATION*. Metadata about the data file itself.
2. *TELESCOPE PARAMETERS*. Information about the LAMOST telescope settings during observation.
3. *OBSERVATION PARAMETERS*. Parameters such as date, time, and exposure details.
4. *SPECTROGRAPH PARAMETERS*. Specifications of the spectrograph used in the observation.
5. *WEATHER CONDITION*. Details of the weather conditions during the observation.
6. *DATA REDUCTION PARAMETERS*. Information about the data reduction process applied.
7. *SPECTRA ANALYSIS RESULTS*. The results from the spectroscopic analysis, including redshifts and classifications.

[4] https://www.lamost.org/dr10/

The FITS files are organized as structured data, containing flux, wavelength, mask, and inverse variance arrays. Low-resolution data involves merging spectra from multiple exposures and stitching together the red and blue arms of the spectrograph.

### *4.2. Medium-resolution Spectroscopic Data Products*

The MRS data products include spectroscopic FITS files and catalog files. Unlike the LRS, MRS does not involve red and blue arm stitching. Instead, both single-exposure and merged spectra are provided, allowing for greater flexibility in data analysis.

## 5. Future

Looking ahead, LAMOST aims to further enhance its data products by expanding the volume of observations and refining data processing techniques. The next data release will include additional spectroscopic observations and incorporate improvements in data reduction algorithms to ensure higher precision and reliability. The future data products will also aim to provide more detailed stellar parameters, along with a more user-friendly format for accessing and utilizing the data, thus supporting a wider range of astrophysical studies.

In addition, LAMOST is actively exploring the use of AI large models for spectroscopic parameter measurement and prediction. So far, promising results have been achieved in this direction. Leveraging the power of machine learning, particularly deep learning models, offers the potential to improve both

the accuracy and efficiency of spectroscopic analysis. In the near future, LAMOST intends to develop a spectroscopic classification and parameter measurement pipeline based on a transformer architecture large model. This approach could revolutionize the way that spectral data are processed, providing an automated, precise, and scalable solution for handling the vast data sets generated by large-scale surveys like LAMOST. These advancements will play a crucial role in advancing the quality and scope of the data products, further enabling groundbreaking research in astrophysics.

## Acknowledgments

This paper was supported by the Young Data Scientist Program of the China National Astronomical Data Center.

## ORCID iDs

Xiao Kong https://orcid.org/0000-0001-8011-8401

## References

Adam, B., Schlegel David, S., Aubourg, J., et al. 2012, AJ, 144, 5
Chen, J. J., Bai, Z. R., Luo, A. L., & Z., Y. H. 2014, SPIE, 9149, 1
Cui, X.-Q., Zhao, Y. H., Chu, Y.-Q., et al. 2012, RAA, 12, 1197
Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., et al. 2016, A&A, 595, A1
Gaia Collaboration, Vallenari, A., Brown, A. G. A., Prusti, T., et al. 2023, A&A, 674, A1
Heather, F. 2018, AAS Meeting Abstracts, 231, 436.01
Hutchinson, T. A., Bolton, A. S., Dawson, K. S., et al. 2016, AJ, 152, 6
Luo, A.-L., Zhang, H.-T., Zhao, Y.-H., et al. 2012, RAA, 12, 1243
Tonry, J. L., Stubbs, C. W., Lykke, K. R., et al. 2012, ApJ, 750, 2
Wei, P., Luo, A. L., Li, Y. B., et al. 2014, AJ, 147, 101
Zhong, J., Li, J., Carlin, J. L., et al. 2019, ApJS, 244, 1