Publication Year	2023
Acceptance in OA	2025-02-28T13:25:49Z
Title	Gaia Data Release 3. Processing and validation of BP/RP low-resolution spectral data
Authors	De Angeli, F., Weiler, M., MONTEGRIFFO, Paolo, Evans, D. W., Riello, M., Andrae, R., Carrasco, J. M., Busso, G., Burgess, P. W., Cacciari, C., Davidson, M., Harrison, D. L., Hodgkin, S. T., Jordi, C., Osborne, P. J., PANCINO, Elena, ALTAVILLA, Giuseppe, Barstow, M. A., Bailer-Jones, C. A. L., BELLAZZINI, Michele, Brown, A. G. A., CASTELLANI, Marco, Cowell, S., Delchambre, L., DE LUISE, Fiore, Diener, C., Fabricius, C., Fouesneau, M., Frémat, Y., Gilmore, G., Giuffrida, G., Hambly, N. C., Hidalgo, S., Holland, G., Kostrzewa-Rutkowska, Z., van Leeuwen, F., Lobel, A., MARINONI, Silvia, Miller, N., Pagani, C., Palaversa, L., PIERSIMONI, Anna Marina, Ragaini, S., RAINER, Monica, Richards, P. J., Rixon, G. T., Ruz-Mieres, D., SANNA, Nicoletta, Sarro, L. M., Rowell, N., SORDO, Rosanna, Walton, N. A., Yoldas, A.
Publisher's version (DOI)	10.1051/0004-6361/202243680
Handle	http://hdl.handle.net/20.500.12386/36331
Journal	ASTRONOMY & ASTROPHYSICS
Volume	674

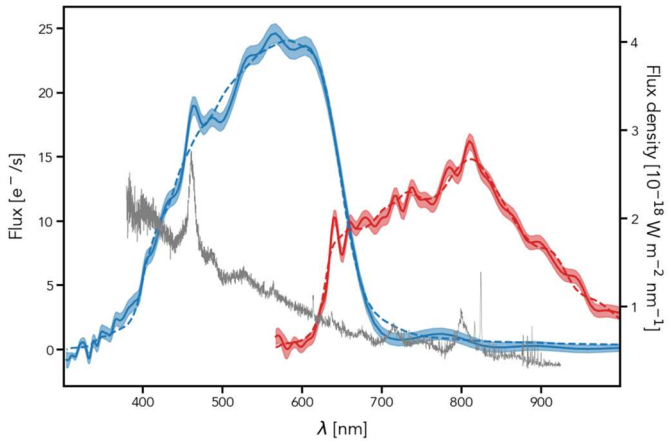

Fig. 17. Comparison between the internally calibrated BP (in blue) and RP (in red) spectra vs. the SDSS (in grey) spectrum for QSO Gaia DR3 5784152373016115 20 (SDSS thing_id = 144680521). Dashed lines are used for the truncated spectra (using only 3 bases for BP and 11 for RP), while continuous lines show the spectra obtained using the full set of 55 coefficients.

this type of stars. For these faint stars, the suppression of noise might aid the data analysis.

The result of the truncation assessment is provided as part of the *Gaia* DR3 in the parameters bp_n_relevant_bases and rp_n_relevant_bases available in the xp_summary table and in the mean continuous spectra available via Datalink (see also Sect. 4). In the case of very faint and typical stars, the use of the truncated representation of BP and RP spectra might be useful. Particularly for sources with unusual spectral energy distributions, such as sources with emission lines, the use of all 55 coefficients for BP and RP, respectively, is advised. The full array of 55 coefficients is available via the archive. Users will need to decide whether or not the suggested truncation is appropriate for their use case.

4. Output data

This Section describes the BP/RP data available via the *Gaia* archive[5]. The exact number of sources with BP/RP mean spectra in the *Gaia* DR3 release is 219 197 643. This list is the result of several selection criteria. Sources with *G*-band magnitude brighter than 17.65 mag and more than 15 CCD transits contributing to the generation of the mean spectra for both BP and RP were automatically selected. The criterion based on the number of transits leads to a (slightly) non-uniform completeness across the sky (see the density sky distribution in Sect. 7). From this initial list, sources that had shown poor estimates of SSC values (see Sect. 8.2 for more details Riello et al. 2021) were excluded unless they were part of one of the lists of specific objects (see below). An additional 35K sources were excluded to allow further processing and validation within DPAC which is likely to be finalised only after *Gaia* DR3. A few lists of specific objects for which other criteria would not apply were defined: these included about 500 sources used for the calibration of the BP/RP data, a catalogue of about 100K WD candidates, 17K galaxies, about 100K quasars, about 19K ultra-cool dwarfs, 900 objects that were considered to be the most representative sources (or centroid) for each of the 900 neurons of the self-organising map used by the Outlier Analysis module (Creevey et al. 2023), and finally 19 solar analogues. All these selections are specific to *Gaia* DR3 and will not affect the content of future releases. In *Gaia* DR3, there is one source (Gaia DR3 5405570973190252288) that has only an RP spectrum.

The gaia_source table in the archive contains a boolean column has_xp_continuous that is true if the corresponding source has BP/RP mean spectra available[6]. After retrieving a list of gaia_source entries, BP/RP spectra can be downloaded from the archive via Datalink[7] in various file formats. This can be done either from the archive web interface or programmatically. In Appendix A we provide instructions for downloading the data from Python.

The spectra are provided in the continuous representation (see also Appendix B for more details): for each BP and RP, the spectrum is defined as a set of coefficients (bp/rp_coefficients); an array with the coefficient formal errors, defined as the standard uncertainties from the least square solution multiplied by the standard deviation of the solution (bp/rp_coefficient_errors); the correlation matrix[8] (bp/rp_coefficient_correlations); various parameters from the source update process, such as number of measurements, number of degrees of freedom, χ^2 and standard deviation of the solution.

In addition to the data available via Datalink, the xp_summary table provides access to some of the parameters listed in the previous paragraph via queries (e.g., to enable the selection of sources based on the standard deviation of their mean spectrum solution) and to other relevant information. Users interested in retrieving the number of CCD transit spectra (and individual measurements) that contributed to the generation of the mean spectrum or that want to know how many of these were assessed as contaminated or blended should interrogate this table, not the main gaia_source table which instead provides similar counters for the photometric data. While BP/RP spectra and *G*-band and BP/RP photometry share part of the processing and filtering criteria, there are also some important differences that can lead to apparent inconsistencies in these counters.

The Python package GaiaXPy[9] has been developed to help the users of BP/RP spectra. It offers the following functionalities: generation of a sampled version of the original continuous representation in both internal and absolute flux and wavelength systems, computation of synthetic photometry in various photometric systems and simulation of *Gaia*-like mean spectra from an input absolute spectral energy distribution. For more information on these tools, we refer to the online package documentation.

5. Validation

5.1. Errors

In order to test the performance of the calibration, a special validation dataset was generated where for each source the available

[5] https://gea.esac.esa.int/archive/

[6] When querying the gaia_source table for sources fulfilling some criteria and having BP/RP spectra available, the user needs to add WHERE has_xp_continuous='true' to the ADQL query.

[7] See https://www.cosmos.esa.int/web/gaia-users/archive/ancillary-data

[8] Given the symmetry of the correlation matrix, only the upper triangular elements (above and not including the diagonal elements which are 1 by definition) of the matrix are provided. The matrix elements are stored as a 1D array of size $n(n-1)/2$ where n is the number of coefficients. The full correlation matrix would therefore be of size $n \times n$. The ordering of the elements in the array follows a column-major scheme.

[9] https://gaia-dpci.github.io/GaiaXPy-website/

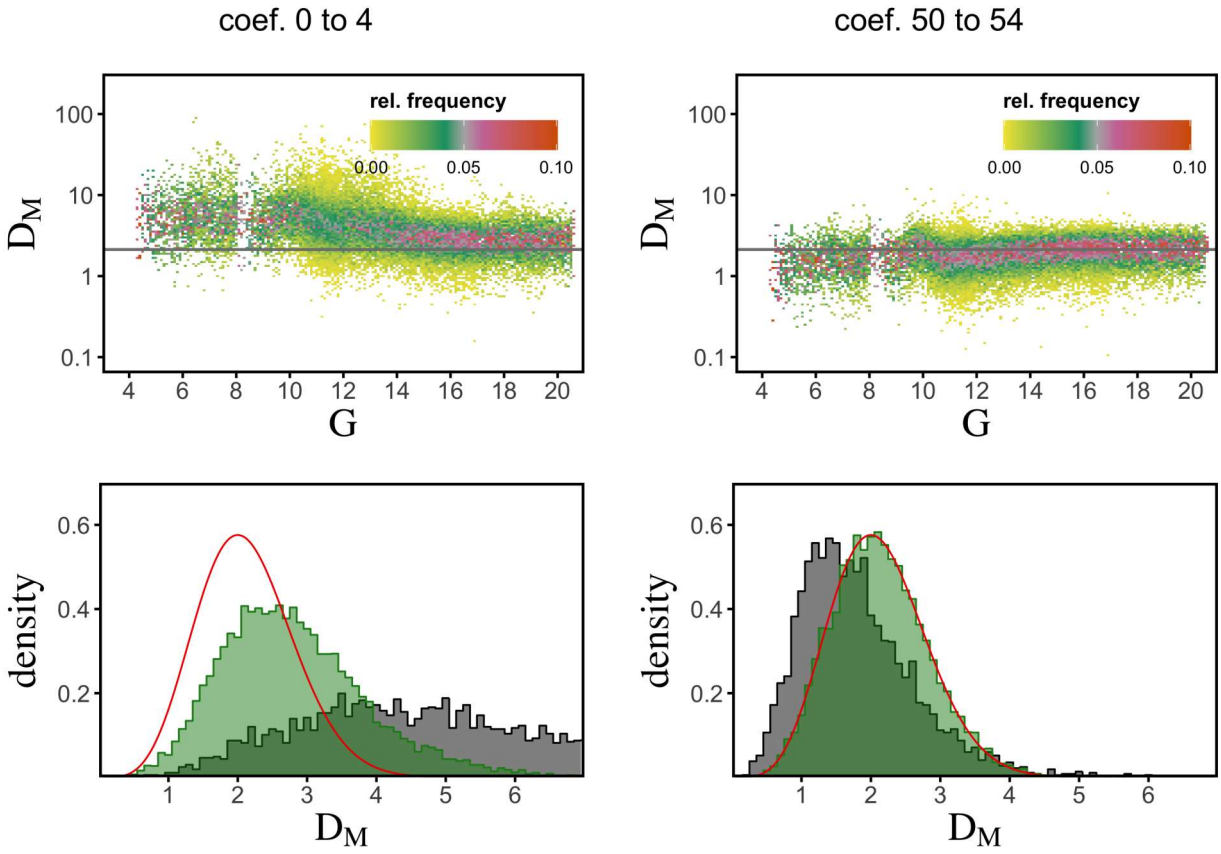

Fig. 18. *Top panels*: distribution of the Mahalanobis distances of all test sources as a function of G-band magnitude. The grey horizontal line indicates the mean of the chi distribution. *Bottom panels*: histograms of the Mahalanobis distances for sources with $G < 10$ mag (grey) and $G > 16$ mag (green). The red line is the corresponding chi distribution. The *left-hand side plots* are for the first five coefficients, with indices 0–4, and the *right-hand side plots* are for the five coefficients of highest order, with indices 50–54.

transits were randomly divided into two groups and processed separately to generate two mean spectra for BP and two for RP. This allows us to compare the calibration results from two sets of transits for the same sources. We refer to this dataset as the BP/RP split-epoch validation dataset. Further details (including how to access the dataset) are available in Appendix D.

For this comparison, we computed the Mahalanobis distance, D_M between the two solutions for each source, given by

$$D_M = \sqrt{(c_1 - c_2)^\mathsf{T} \, (\Sigma_1 + \Sigma_2)^{-1} \, (c_1 - c_2)} \, . \tag{5}$$

Here, c_1 and c_2 denote the coefficient vectors for the two solutions, and Σ_1 and Σ_2 the corresponding covariance matrices. Under the idealised circumstances of normally distributed noise, correct covariance matrices, and the absence of intrinsic photometric variability of the sources used in the test, D_M follows a chi distribution with the degree of freedom corresponding to the length of c_1 and c_2. Deviations from a chi distribution therefore indicate unreliable covariance matrices Σ_1 or Σ_2.

We analysed the distribution of the D_M in comparison to the chi distribution as a function of colour, magnitude, and indices of coefficients. The dependency on colour is only weak, with slightly larger values of D_M for very red sources, with $G_{\rm BP} - G_{\rm RP} \gtrsim 3.0$ mag. The magnitude dependency is more pronounced, and depends on the indices of the coefficients. This is illustrated in Fig. 18. The top panels of this figure show the distribution of the D_M, normalised to the total number of sources in each magnitude bin, for all 40K test sources, for the first five and the last five coefficients in BP, respectively. For the first five coefficients, the values of D_M are in general too large compared to what is expected from a chi distribution, an effect that is more pronounced for bright sources. For the five coefficients corresponding to the highest order basis functions, the magnitude dependency is weaker, with values being slightly smaller than expected from a chi distribution for the brighter sources.

The bottom panels of Fig. 18 show the density histograms for bright sources, with $G < 10$ mag in grey, and faint sources, with $G > 16$ mag in green, respectively, in comparison with the chi distribution for five degrees of freedom. For the first five coefficients, the distribution is much wider than the chi distribution, in particular for the bright sources, and is shifted to larger values. For the last five coefficients, the faint sources are in good agreement with a chi distribution, while the distribution for the bright sources is shifted towards smaller values of D_M.

An underestimation of the error results in larger D_M than expected from a chi distribution, while an overestimation of the error results in smaller values. The differences in D_M with respect to the chi distribution can therefore be interpreted as an underestimation of the errors for the coefficients with low indices, and an overestimation of the errors for coefficients of high indices for bright sources. For high indices and faint sources, the errors are however reliable. While the results shown here are from BP spectra, the situation for RP is similar. When using the BP/RP spectra, the errors for brighter sources in particular should be interpreted with caution.

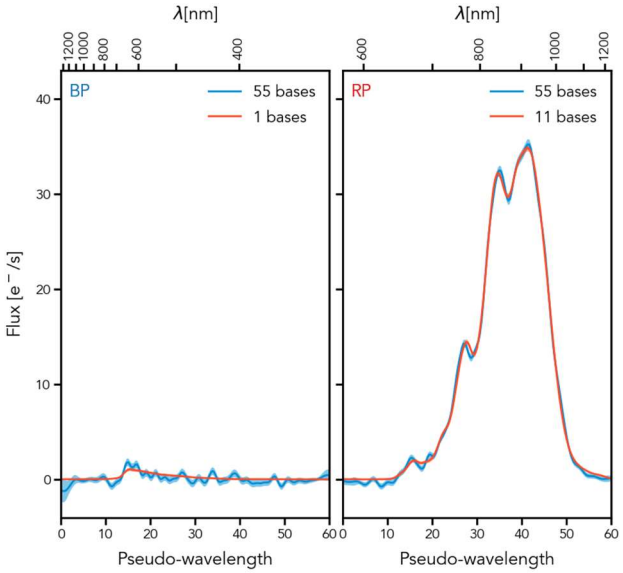

Fig. 19. BP (*left*) and RP (*right*) spectra for the faint red source Gaia DR3 1252666141462905344 (G_{BP} = 21.6 mag and G_{RP} = 17.8 mag). The blue curves show the spectra defined by the 55 coefficients (errors are shown as a shaded area). The red curves show the truncated spectra where only the first bp/rp_n_relevant_bases have been used.

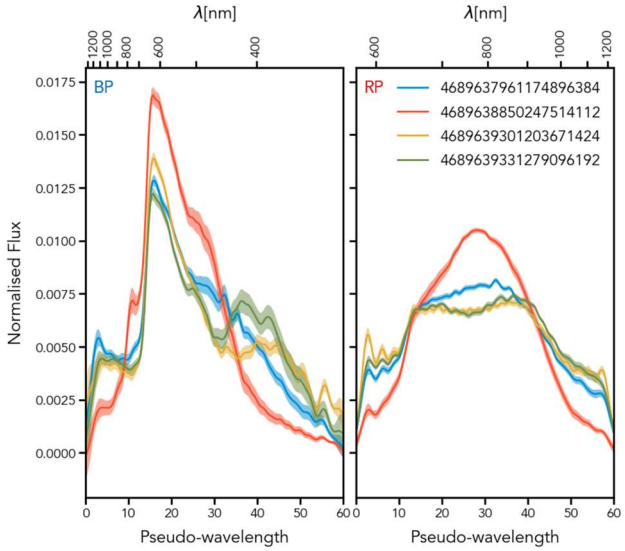

Fig. 20. BP (*left*) and RP (*right*) normalised internal spectra of some sources with all transits blended by other nearby sources in the 47 Tuc cluster.

5.2. Specific cases

Although most of the spectra show a good behaviour, there are a few cases where we see peculiar shapes, which are due to several factors. In the following, we analyse a few of the most common situations.

In the case of very faint sources, the fitting procedure generating the mean spectrum will be poorly constrained and may produce unrealistic features. For example, Fig. 19 shows the spectra of a faint red source (with G_{BP} = 21.6 mag and G_{RP} = 17.8 mag). For this type of spectrum, the parameters bp_n_relevant_bases and rp_n_relevant_bases in the xp_summary table in the *Gaia* DR3 archive are particularly relevant, as they indicate the number of coefficients that are significant considering the noise level (see Carrasco et al. 2021, and Sect. 3.4.3 in this paper for more details). In this case, only 1 of the 55 coefficients defining the BP spectrum is considered significant. Our adopted truncation procedure suggests that, for BP, all coefficients beyond the first one are only fitting the noise fluctuations rather than real spectral features and can be ignored when using the mean spectra for further investigations. For RP, the number increases to 11 thanks to the higher S/N.

In crowded areas, it is possible that two or more sources are so close in the sky that their observations are always or often contaminated or blended. We refer to blended spectra when two or more sources fall within the observed window, while contamination refers to flux belonging to a source that is located outside the window. If this happens in a large fraction of the observations of a given source, then the mean spectra for that source will be affected. To enable users to assess the reliability of BP/RP mean spectra, the xp_summary table in the archive includes several parameters (bp/rp_n_blended_transits and bp/rp_n_contaminated_transits) indicating the number of transits affected by blending or contamination for all sources for which BP/RP spectra are published. Figure 20 shows the case of four sources in the globular cluster 47 Tuc that have all their observations flagged as blended. Users are strongly encouraged to make use of the available crowding flags to detect problematic cases.

The wings of the spectra should normally have a low flux level because of the combined action of LSF, dispersion, and response. However, this may not be the case because of the presence of residual background flux not fully removed in the background calibration stage or diffused flux due to the source being extended. For example, Fig. 21 shows the BP and RP internal spectra for a source with the in_galaxy_candidates flag in the gaia_source table set to true. Both spectra present a higher-than-normal flux in the wings. This source also shows a significant mismatch between the photometry in the different bands (G = 18.7 mag, G_{BP} = 15.7 mag and G_{RP} = 14.3 mag), the two BP/RP integrated flux values being much brighter than the value in the G-band, due to the much larger size of the BP/RP windows with respect to the AF ones.

A similar effect is seen when considering objects that are close to a very bright source. Their spectra will appear to be contaminated by flux coming from the nearby bright object. The resolution of the background calibration is not sufficient to completely remove this effect and may actually lead to an under- or overestimation of the background in the regions surrounding very bright sources. Figure 22 shows the BP/RP spectra for two sources near Sirius. Source Gaia DR3 2947050466531872640, at 30 arcsec from Sirius, is clearly contaminated by diffuse flux coming from the nearby bright source. Also in this case, the photometry indicates a much brighter source in the BP/RP integrated bands than in the G-band: G = 15.7 mag, G_{BP} = 13.2 mag, and G_{RP} = 13.2 mag. The second source (Gaia DR3 2947047202356748672) is located further away at about 3 arcmin. In this case, the background seems to have been overestimated, causing negative flux values in the wing of the spectra in both BP and RP.

5.3. Signal-to-noise ratio

An overall indication of the S/N for a given source and photometer can be obtained directly from the coefficients by dividing the L2-norm of the vector of coefficients by the L2-norm

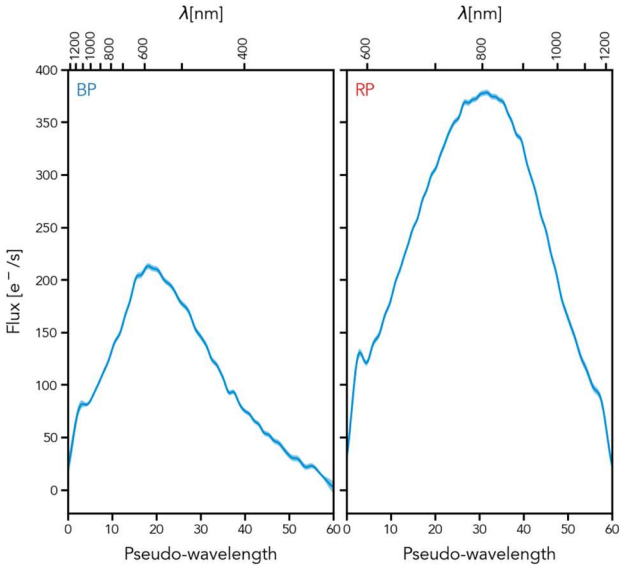

Fig. 21. BP (*left*) and RP (*right*) internally calibrated spectra of a source (Gaia DR3 1252344813484742272) flagged as `galaxy` in the `gaia_source` table. The spectra are broader than expected and the corresponding integrated magnitudes are much brighter compared with the *G*-band photometry.

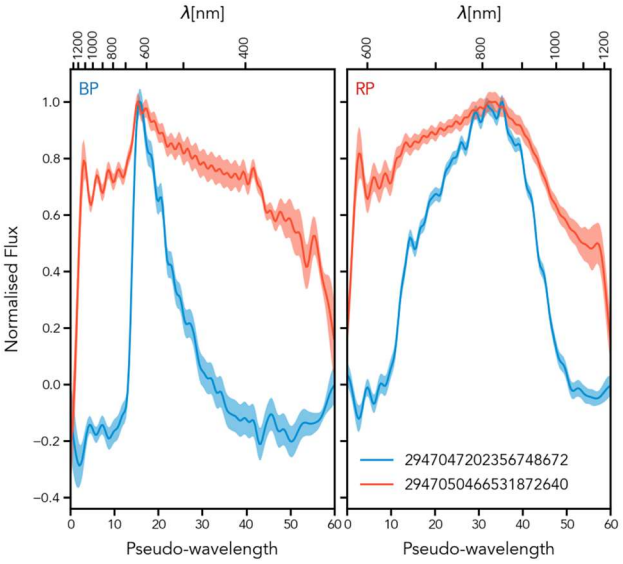

Fig. 22. BP (*left*) and RP (*right*) internally calibrated spectra of two sources near Sirius: one located at 30 arcsec (in red) and the other at 3 arcmin (in blue). The source closest to Sirius shows clear signs of contamination from the nearby object.

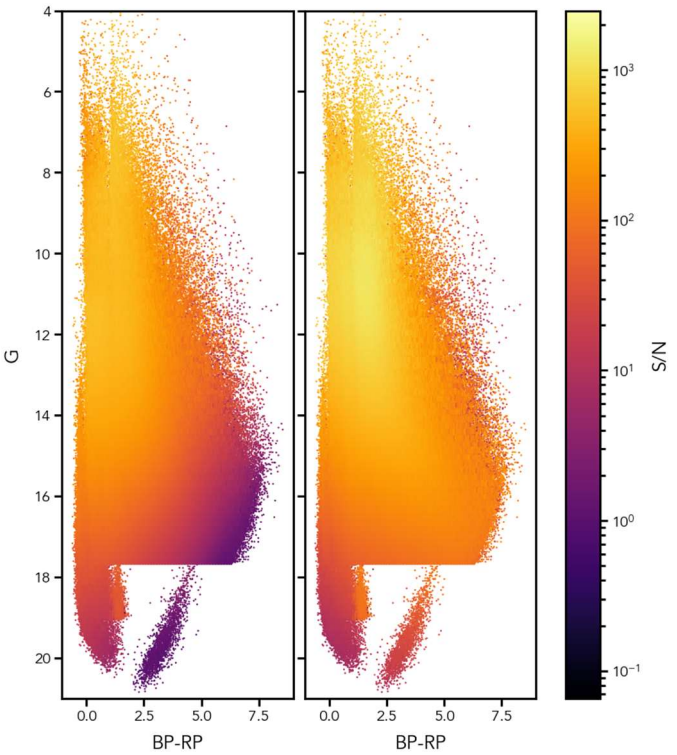

Fig. 23. Colour–magnitude diagram of a random 10% of the sources for which BP/RP spectra are available in *Gaia* DR3, colour coded in a logarithmic scale by the global S/N as computed directly from the continuous representation coefficients and their errors. BP and RP S/Ns are shown in the *left and right panels*, respectively.

of the vector of errors on the coefficients. Figure 23 shows a colour–magnitude diagram of the sources with BP/RP spectra in *Gaia* DR3 colour coded by this global S/N in the BP and RP photometers in the left and right panel.

A user that is interested in the S/N at different wavelengths will have to consider the representation of the spectrum by the linear combination of basis functions that have an explicit wavelength dependency rather than relying on the coefficients alone. The panels in Fig. 24 show typical S/N distribution of internally calibrated spectra over the BP (left panels) and RP (right panels) pseudo-wavelength ranges covered by the BP/RP spectra. In the top two panels, each curve shows the S/N for sources of different magnitude, as reported in the colour bar, $G_{\rm BP} - G_{\rm RP}$ colour close to 1.0, and with typical global S/N (for sources of similar magnitude and colour). In the bottom two panels instead, each curve shows the S/N for sources of different colour, as reported in the colour bar, *G*-band magnitude close to 16.0 and with typical global S/N (for sources of similar magnitude and colour). Only sources with $|c_*| < 0.02$ have been considered for these plots, c_* being the corrected BP/RP flux excess factor as defined in Riello et al. (2021). As in previous figures, the top axes showing the correspondence with absolute wavelengths are only indicative.

Due to the fact that the mean BP/RP spectra are a combination of many single observations for each object, intrinsic variability will result in larger uncertainties in the mean spectra. This is confirmed by the fact that the S/N for a sample of known RR Lyrae (extracted from Clementini et al. 2023) is significantly lower than the S/N for a sample of random (mostly non-variable) sources with similar apparent *G*.

The dependency of the S/N from pseudo-wavelength is linked to the spectrum itself. Looking at the top-right panel of Fig. 24, the maximum S/N in RP is achieved for sources with *G* 9–10. Saturation and occasional gate misconfiguration could be responsible for this: while the mean spectra of very bright sources do not show clear signatures of saturation, the presence of some saturated epoch spectra among those contributing to the mean spectrum –which is possibly due to gate misconfiguration caused by large on-board magnitude errors at the bright end– could lead to a larger scatter around the peak and therefore a larger error and a lower S/N than expected.

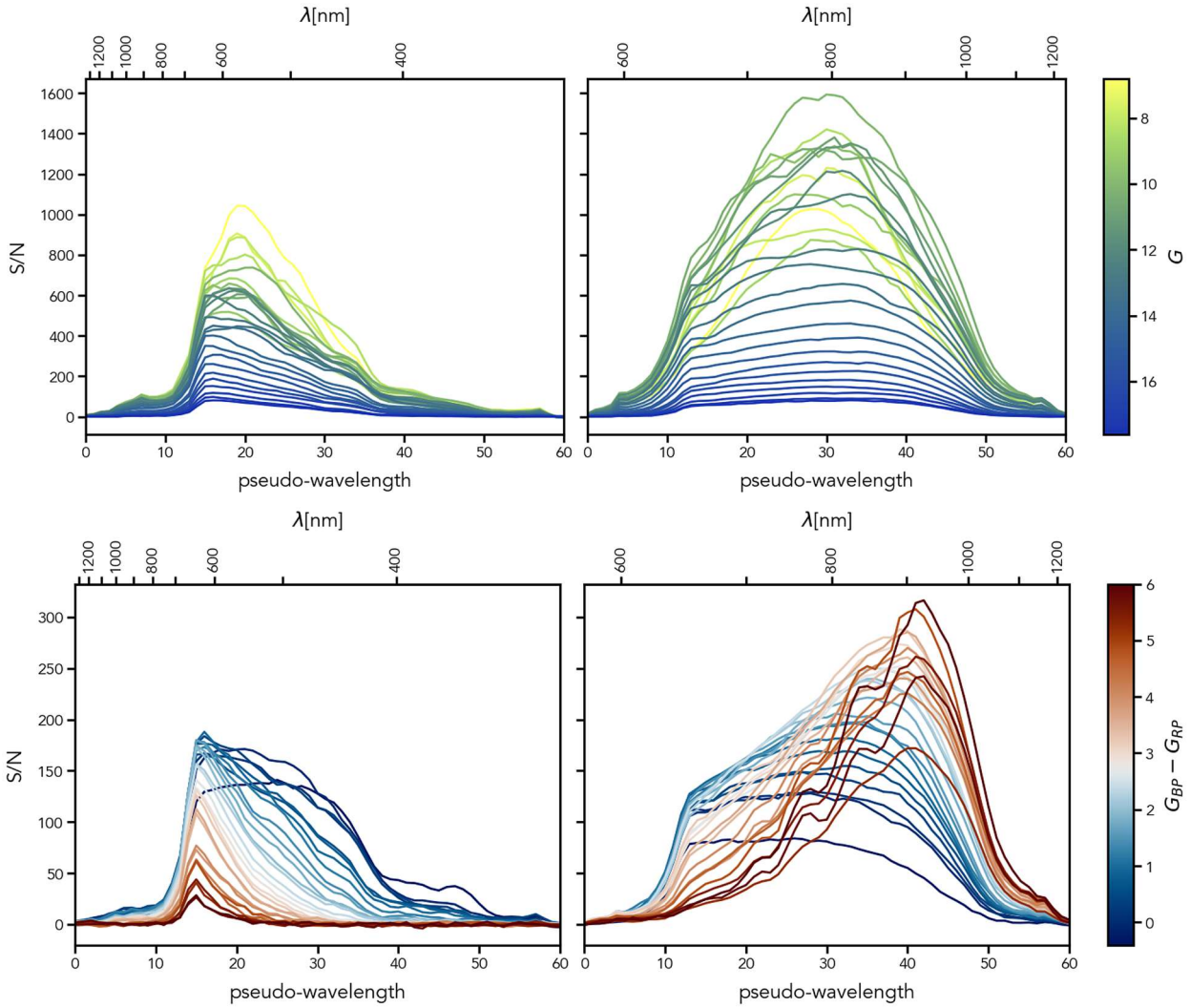

Fig. 24. S/N vs. pseudo-wavelength (and approximate absolute wavelength) for internally calibrated spectra. The *top panels* show the S/N for sources of different magnitude and similar colour (close to 1.0), while the *bottom panels* focus on sources with similar *G*-band magnitude (close to 16.0) and a range of colours.

6. Recommendations

6.1. Recommended format

The mean spectra are available in the archive in the form of a set of coefficients that define a continuous function over the pseudo-wavelength range. This is the fundamental product of the BP/RP spectral data processing. When sampling the spectra on a discrete grid in pseudo-wavelength (or wavelength if working in the absolute system), some information is unavoidably lost. In particular, the continuous representation comes with full covariance information, whereas a spectrum sampled on a (pseudo-)wavelength grid with more points than the number of coefficients in the continuous representation cannot. Users are therefore strongly encouraged to consider using the continuous representation to best exploit the BP/RP spectra in *Gaia* DR3 (e.g., to derive astrophysical parameters or analyse the presence of spectral features) and avoid sampling the spectra or deriving synthetic photometry from them, losing information in the process.

Figure 25 shows that the coefficients can be used to successfully classify sources in different regions of the Hertzsprung–Russell diagram. At least for the few cases shown in the plot, most of the information required for classification is already available in the first few coefficients of the continuous representation. Figure 26 shows the corresponding plot with the more familiar sampled spectra.

Figure 13 clearly shows that narrow spectral features in the spectra can only be reproduced with larger higher order coefficients. For example, Fig. 27 shows an example of two sources with rather similar RP spectra except for the presence of a strong emission line. One of the two sources is a QSO. As can be seen in the bottom right panel, higher order coefficients for the QSO have larger values.

6.2. Effects of noise

The correlations between the coefficients of a source, both for BP and RP, are in general rather low, with median correlation coefficients well below 0.1 in both BP and RP. When constructing the sampled spectrum as a function of pseudo-wavelength (or wavelength), the correlations might become much more important. As there are only 55 basis functions for BP and RP, respectively, any sampled spectrum with more than 55 sample points needs to have linear dependencies among the samples. Furthermore, even

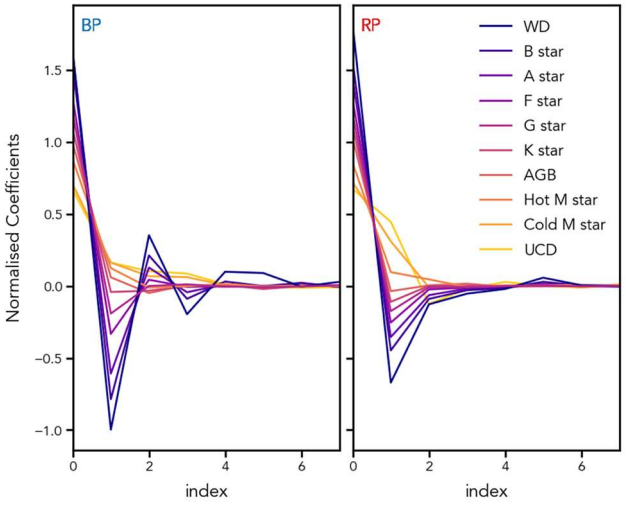

Fig. 25. First eight coefficients of the continuous representation in BP (*left*) and RP (*right*) for some sources with different astrophysical parameters.

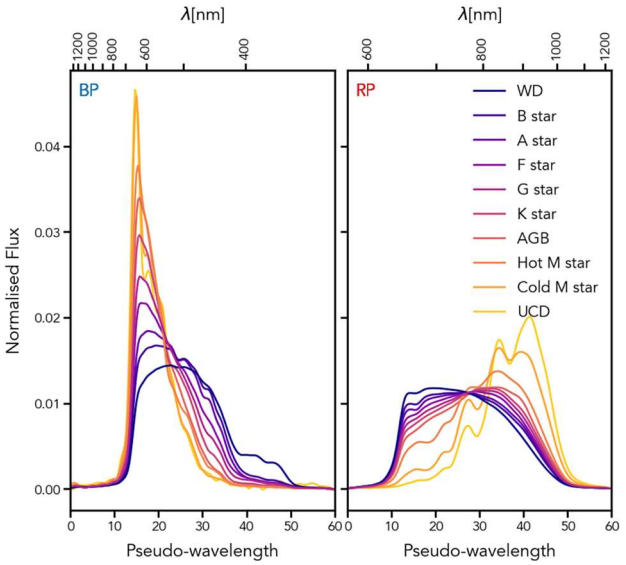

Fig. 26. Normalised internal mean spectra in BP (*left*) and RP (*right*) for the same sources shown in Fig. 25.

if the coefficients were uncorrelated, the non-local character of the basis function representation would still introduce correlations between different pseudo-wavelengths. This effect is illustrated in Fig. 28 for the RP spectrum of one particular source, with $G = 17.89$, $G_{\rm BP} - G_{\rm RP} = 2.74$. The BP/RP split epoch validation dataset (see Appendix D) has been used for this analysis. The two sets of transits for this source contain 18 transits and 3 transits, respectively. Consequently, the S/N in the first set is higher than in the second one. This is seen in the first column of Fig. 28, where the coefficients for the calibration using only three transits are noisier and have larger error bars than for the 18 transits case. The second column in this figure shows the correlation matrices for the two cases. In general the correlations are low, with little structure in the off-diagonal entries. However, the correlations are larger for the noisier case. The third column shows the sampled RP spectra for the 18 transits and the three transits cases. The larger noise in the latter case manifests itself in a wavy structure in the sampled spectrum. In the correlation matrix for

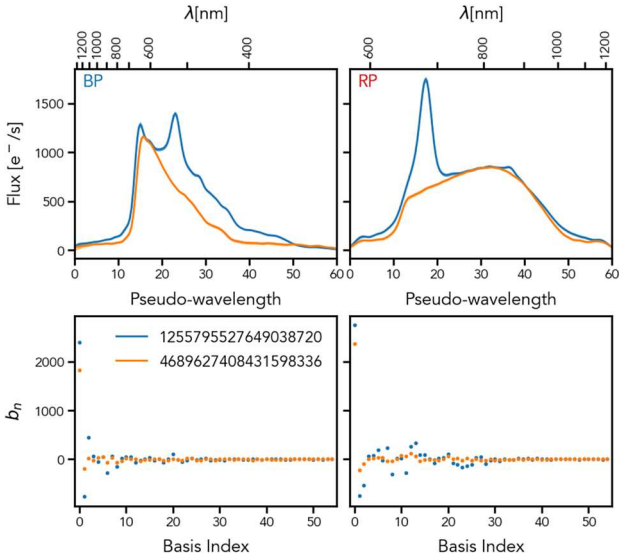

Fig. 27. Comparison of the mean spectra obtained for a QSO with a strong emission line (`Gaia DR3 1255795527649038720` in blue), and another source with similar shape and flux level but without strong features (`Gaia DR3 4689627408431598336` in orange). BP and RP are shown in the *left and right panels*, respectively. Sampled spectra are shown in the top panels, while the bottom panels show the corresponding coefficients.

the sampled spectrum, shown in the fourth column, this larger noise manifests itself in the form of alternating short-scale patterns of positive and negative correlations. These patterns are again more pronounced when the S/N is lower. As random noise in the BP/RP spectra manifests itself in the sampled spectra as wavy structures, and correlations within the sampled spectra are not negligible, the interpretation of the coefficients, being much less affected by correlations, might be more convenient.

7. Conclusions

In this paper, we focus on the processing that generated the internally calibrated BP/RP spectra contributing to *Gaia* DR3 starting from the raw satellite data. The released data are time-averaged source spectra that result from the combination of all single observations of a given source. Only a selection of all generated spectra will be included in the release at this stage, but several other new products are based on the entire dataset. The main challenges faced by this step in the data processing are due to the vast amount of data (about 65 billion single BP/RP transits were processed), to the nature of the low-resolution aperture prism spectroscopy with the additional complications added by the TDI mode, and to the large number of different observing configurations effectively corresponding to the different instruments that need to be calibrated onto the same homogeneous system. We explain how we dealt with these challenges and show how we have been monitoring the intermediate performances of our calibration procedures. We also describe the somewhat unfamiliar format of the BP/RP spectral data in the archive. Rather than providing spectra defined as a flux value corresponding to a sample covering a given wavelength range, the BP/RP spectra are represented by an array of coefficients, and their errors and correlations, which are to be applied to a set of basis functions to obtain a continuous function. This approach allows us to combine multiple transit spectra, each having its own sampling, dispersion, and LSF (Carrasco et al. 2021). The set of

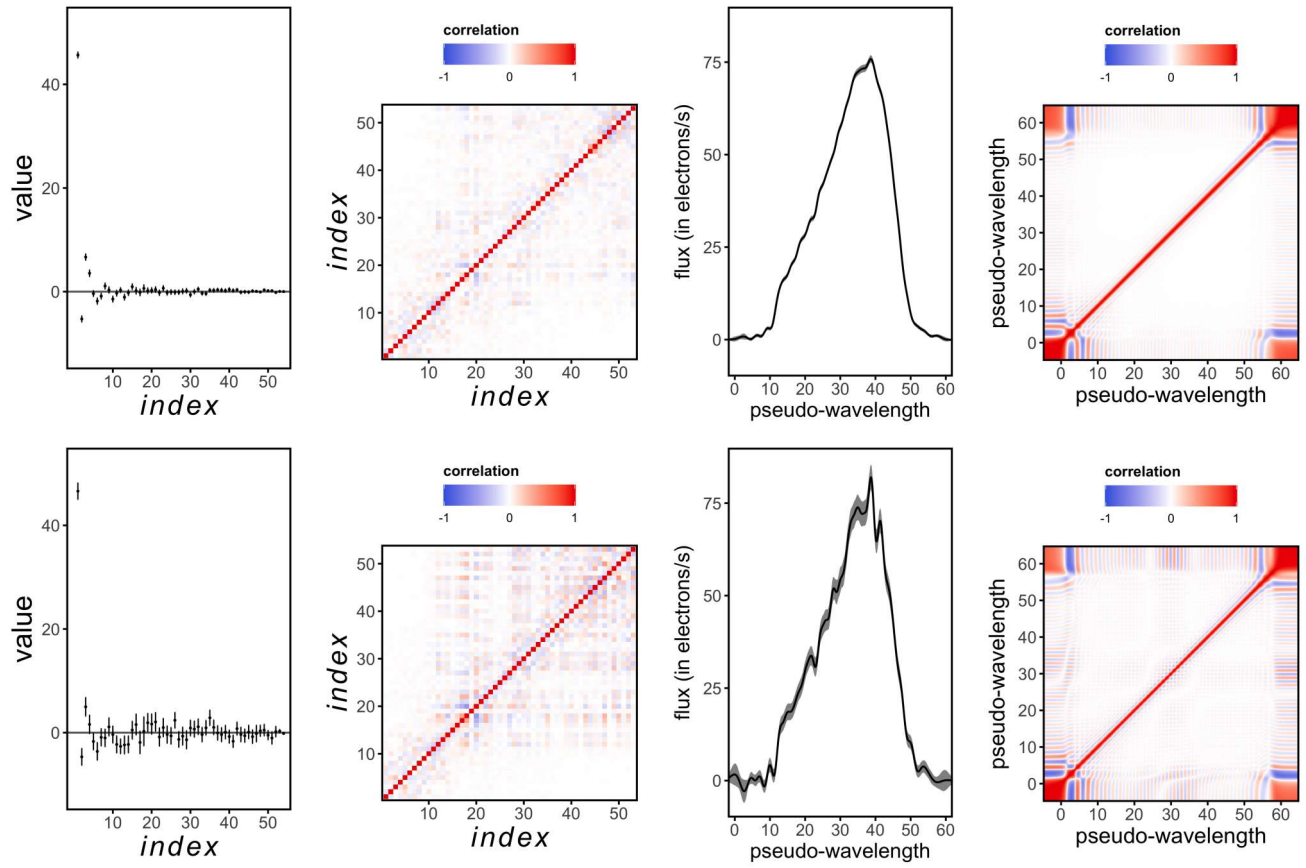

Fig. 28. Example of the effect of noise for an RP spectrum. *First column*: RP coefficients with errors. *Second column*: correlation matrices for the coefficients. *Third column*: sampled RP spectrum (black line) with 1-sigma uncertainty interval (grey shaded region). *Fourth column*: correlation matrix for the sampled RP spectrum. The *top row* is for 18 transits, and the *bottom row* for 3 transits, for the same source.

bases has been optimised to ensure maximum efficiency, thus focusing most of the flux in the first few coefficients and leaving higher order coefficients to be constrained by narrow spectral features.

We want to conclude this paper by showing some sky distributions related to the BP/RP data in Fig. 29. All maps are in Galactic coordinates and show the entire catalogue of sources with BP/RP spectra in *Gaia* DR3. The first map shows the density distribution in the sky. As expected, most of the sources are concentrated along the Galactic plane. The two Magellanic Clouds also stand out, as well as a few clusters. The darkest areas close to the Galactic plane in the map correspond to regions obscured by dust and regions with extremely high density where the BP/RP data are particularly affected by strong crowding (both in the acquisition and in the processing). Some regions with lower density away from the Galactic plane still show imprints of the scanning law (compare this with the map showing the median number of transits). These are expected to disappear with the addition of more observations in future releases. The second map shows the distribution of $G_{\rm BP} - G_{\rm RP}$ colour. The third map shows the median number of transits per source (in RP). This is clearly defined by the satellite scanning law. A similar map of BP would be very similar with the exception of the occurrences of a larger number of transits near the Ecliptic poles. These are due to the first month of operations in Ecliptic scanning law. This period was not included in the generation of average source BP spectra as explained in Sect. 3.3. The fourth map shows the median fraction of contaminated or blended transits with respect to the number of transits per source for RP. The equivalent maps for BP would be very similar. The areas showing higher density in the first map also stand out in this map as regions where the mean spectra are more affected by crowding. This is justified by the fact that the crowding evaluation is limited to the *Gaia* source catalogue itself. Finally, the last two maps show the distribution in the sky of the median of the 84th percentile of the S/N distribution over the BP and RP wavelength ranges. As expected, the scanning law signature is very evident in these maps with errors being lower in the most observed regions. Areas at low Galactic latitude show lower S/N in the BP spectra due to the abundance of red-coloured sources. The S/N distribution of the internally calibrated spectra shows values larger than 1000 for bright sources in some wavelength ranges (see Fig. 24). *Gaia* DR3 will contain about 700 000 BP spectra and 4.3 million RP spectra with the 84th percentile of the S/N above 500.

Various parameters available from the archive can be useful to clean the catalogue from disturbed spectra. A very useful quantity already introduced for *Gaia* DR2 is the `phot_bp_rp_excess_factor`. This parameter is available from the `gaia_source` table and is defined as the ratio between the sum of BP and RP integrated fluxes and the G-band flux for the same source. Due to the shape of the G, $G_{\rm BP}$, and $G_{\rm RP}$ passbands, some colour dependency of this ratio is expected and may bias selections based on `phot_bp_rp_excess_factor`. To correct for the expected colour trends, users should apply the equation recommended in Riello et al. (2021) to form what is

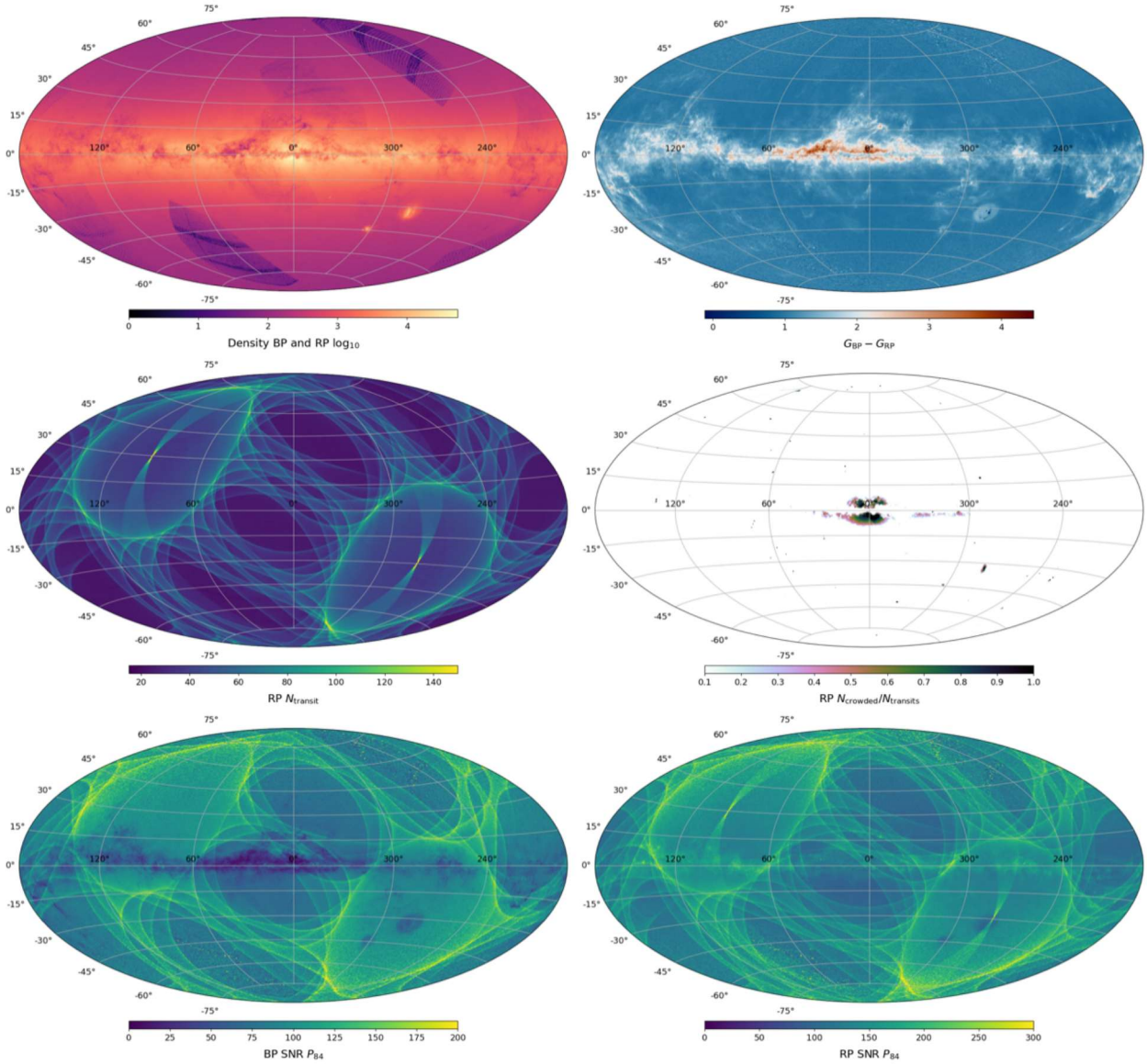

Fig. 29. Sky distribution (in Galactic coordinates in Hammer-Aitoff projection, with resolution equivalent to HEALPiX level 7) of various parameters related to the BP/RP data: *from the top left to the bottom right* the maps show the sky density of objects with BP/RP spectral data, the median $G_{BP} - G_{RP}$ colour, the median number of transits in RP contributing to the mean spectra, the median crowding level, and the median of the 84th percentile of the S/N over the BP and RP ranges. The colour scales do not cover the full range covered by the data.

known as $C^\star$[10]. The deviation of this parameter from 0.0 indicates the presence of inconsistencies between the flux measured in the BP/RP windows and the flux in the G-band. These inconsistencies can be due to different source properties (e.g., in the case of extended sources) or systematic errors in the calibration procedures (e.g., in the case of residual background due to nearby bright sources). Section 9.4 in Riello et al. (2021) also provides a function reproducing the 1σ scatter for a sample of well-behaved isolated stellar sources with good-quality photometry. Users wishing to use $C^\star$ and its 1σ scatter to select the most reliable spectra would find that 90% of the sources have $C^\star < 3\sigma$ while 79% fulfil the criterion $C^\star < 1\sigma$. Figure 30 shows the distribution of $C^\star$ together with the 1- and 3-σ limits.

In terms of BP/RP spectral data, future releases will see a vast increase in the number of average source spectra and the addition of calibrated epoch spectra, that is, spectra derived from one single observation in BP/RP. From a processing and validation point of view, this will focus the attention on calibrations that deviate from the average behaviour. While robust techniques help mitigate these problems when generating mean spectra, the application of noisy calibrations can generate unreliable data. This needs to be mitigated to ensure the quality of calibrated BP/RP epoch spectra, which we plan to include in future releases. One other area where some improvement is being

[10] $C^\star$ is obtained from the `phot_bp_rp_excess_factor` C as $C - f(G_{BP} - G_{RP})$ where $f(G_{BP} - G_{RP})$ is a polynomial in colour defined as

$$f(x) = \begin{cases} 1.154360 + 0.033772\,x + 0.032277\,x^2 & \text{for } x < 0.5 \\ 1.162004 + 0.011464\,x + 0.049255\,x^2 - & \\ \qquad 0.005879\,x^3 & \text{for } 0.5 \le x < 4.0 \\ 1.057572 + 0.140537\,x & \text{for } x \ge 4.0 \end{cases}$$

where $x = G_{BP} - G_{RP}$.
The corrected parameter (`c_star`) will be available for all sources included in the *Gaia* Synthetic Photometric Catalogue from the archive; see Gaia Collaboration (2023c).