Investigation of possibilities and limitations for bulk automated analysis of big datasets of reflectance spectra of asteroids

Jakub Morawski

January 12, 2024

Abstract

The report describes an attempt to create an optimally robust framework for accessing and analyzing reflectance spectroscopy data from spectroscopic surveys of asteroids. The objective is to be able to download arbitrarily (at least within the order of magnitude comparable to the number of known asteroids) large datasets from a variety of sources and apply functions which can extract scientifically relevant information from an analysis of these spectra. Feasibility of such endeavor is being investigated by exploring six different databases: two smaller datasets corresponding to narrow surveys designed with a specific focus (Fornasier, Gartrelle) and four larger datasets corresponding to broader surveys (PRIMASS-L,SMASS, SMASS II, IRTF). Focus is placed on essential tasks in asteroid spectra analysis: taxonomic classification, absorption band identification and analysis, mineralogical content estimations (for the IRTF dataset only, due to numerous limitations discussed). The code developed for the purpose of this project is published. The study is summarized in terms of major achievements and setbacks, and possibilities to extend the program so as to implement advanced spectra modelling techniques are discussed.

1 Introduction

The last few decades have brought numerous breakthroughs in the study of the Solar System in general, in particular turning researcher's attention more and more towards the smaller rocky bodies. As stated in the book *Asteroids IV* (Michel, DeMeo, and W. F. Bottke 2015), the scope of knowledge acquired from studying asteroids has grown immensely since *Asteroids III* (William F. Bottke Jr. *et al.* 2002). Large surveys on Near Earth Objects (NEOs), such as NEOWISE, have allowed for reaching a Spaceguard goal of detecting 90% of all Near Earth Asteroids (NEAs) above the size of 1 km (Mainzer *et al.* 2011). Paired with simulations and theoretical models, the studies of asteroids' orbits provide insights in the past of the Solar System's evolution and the role of small rocky bodies.

In parallel, an effort is made to analyse the physical and compositional properties of asteroids. A major step forward in that matter has been brought by space missions approaching asteroids at close distance, with Hayabusa (Yano *et al.* 2006) and OSIRIS-REx (Lauretta *et al.* 2017) bringing samples of asteroid material back to Earth (from S type asteroid Itokawa and B type asteroid Bennu, respectively), and Psyche mission on it's way to explore an M-type asteroid (NASA n.d.(a)). Nevertheless, remote sensing remains the primary source of information for the vast a majority of asteroids. In that regard, spectroscopy provides a glimpse into the structure and mineralogical composition. As mentioned in the supplementary report on theoretical models for reflectance spectra (Morawski 2023), we are able to disband solar light reflected from surfaces of planetary bodies, asteroids in particular, the most common and reliable application being a study of absorption bands, with empirical formulas (supported by studies of meteorites) for percentage contents of olivines and pyroxenes dependent on positions and depths of absorption bands around 1μm and 2μm (e. g. Cloutis *et al.* 1986). Spectroscopy has also been used to classify asteroids into spectral classes, or *taxonomies*, with a classification scheme having evolved from Tholen 1984 up to a currently utilized scheme from S. J. Bus *et al.* 2008. Theoreticians strive to create mathematical models for the spectral curves and their dependence on properties of the reflecting medium (rock/regolith on planetary surfaces), two leading ones being the Hapke model (introduced in Hapke 1981, refined in successive papers, most notably Hapke 2002) and the Shkuratov model Shkuratov *et al.* 1999 (see overview in Morawski 2023).

In the regard of remote spectroscopic observations, the most significant step up of the last decades being the increase of the size of available datasets, with dedicated spectroscopic surveys obtaining reflectance spectra of hundreds of asteroids (see section 2). In upcoming future, we can imagine an exponential increase of these datasets. As of the

end of 2023, there have already been 33955 NEAs, and a total of 1329548 minor planetary bodies detected (Union n.d., as of 31.12.2023). If we consider the past of the research of asteroids, we can see that a gradual paradigm shift from an analysis of individual objects to a bulk analysis of large datasets has transpired in the second half of the 20^{th} century. As summarized in DeMeo *et al.* 2015a, the total number of asteroids discovered had amounted to several in the period of $1800 - 1850$ (Ceres being the first asteroid ever discovered, in the year 1801, by an Italian priest and astronomer, Giuseppe Piazzi (NASA n.d.(b))), increased significantly in the second half of the 19^{th} century, but stayed in the order of 10^3 until mid 20^{th} century. From then on, the curve of the number of known asteroids over time gained a much higher slope, the discoveries being made thanks to large astrometric surveys (Palomar-Leiden survey (van Houten-Groeneveld *et al.* 1989), Spacewatch (McMillan 2000), LINEAR (Stokes *et al.* 1998)), as well as, most recently, with *citizen science* (such as "Find an asteroid" campaign of The International Astronomical Search Collaboration, Miller 2016). Accordingly, the methods used to analyse the data needed to be changed - in the early days it was possible to exhaustively study each individual asteroid and determine the orbit with manual calculations. Nowadays, to respond to the needs of increased datasets, and take advantage of technological advances, a lot of the process is automated, with astrometric data being parsed by computer programs, and information on the orbits of asteroids being stored in large databases (N. JetPropulsionLaboratory n.d.(a)). While it can still be valuable to look at specific cases, the bulk of data and automation of analysis techniques allows us to look at the structure of the Solar System as a whole.

If we compare that stage of astrometric research with the current stage of spectroscopic research, we could conclude that the latter is a step behind, with a similar shift being expected, and desired, in years to come. The number of asteroids with spectra available at PlanetaryDataSystem n.d. is of the order of $10^3 - 10^4$, and a lot of data analysis is focused on individual cases. With the continued interest in the subject, and millions of asteroids known to humankind, the numbers will continue growing, leveraging spectroscopy of asteroids to the realm of *big data*. Accordingly, methods to analyse asteroid spectra in bulk need to be developed. One intuitive approach is to apply techniques of machine learning, as has been undertaken in Dyar *et al.* 2023. Nevertheless, an ultimate goal should be implementing same methods of spectra analysis we use for individual cases into algorithms that can be run in a loop for all objects in a database with as minimal a degree of human supervision as possible.

The goal of this work is to study the feasibility of automating different aspects of data analysis for reflectance spectra of asteroids. For this purpose, a *Python* code has been developed to download and analyse the data. A Python package for the analysis of asteroid spectroscopic data, called *CANA* (De Pra *et al.* 2018), has been implemented for some of the tasks. This report describes each stage of the work and looks into perspectives of how these methodologies could be used in scientific research.

Section 2 describes the kind of data available based on the examples that were looked into during this project. In 3, the methodology used for downloading data from different databases (corresponding to different spectroscopic surveys), and the challenges involved relating to differrent data formats, are presented. Section 10 touches on the topic of searching for overlaps between databases (spectra of the same asteroid taken by differrent surveys, or repeatedly within the same survey) and some insights that can be gained from looking at those. Finally, in 11, perspectives are shed on how the code could continue to be developed and yield more results of scientific value.

2 Databases taken under study

The bulk of data files with reflectance spectra of asteroids from different surveys can be accessed at PlanetaryDataSystem n.d. The website provides links to multiple databases and files from various spectroscopic surveys. The general format of a single data file is an arrangement of 3 columns, so that each row has a wavelength value, corresponding relative reflectance and and error on the latter. In some cases, the last column with the error is absent. Exact conventions for names of data files differ from one database to another, but they always include the asteroid number (as assigned by the *International Astronomical Union*) and in many cases also the asteroid name. For the purpose of this project, several databases, described in the following subsections, were looked into. The choice of databases was partially coincidential, largely based on intuition, but also supported by a desire to both seek novelty but also compare with results from well established and acknowledged research. Ultimately, the (largely achieved) goal was to facilitate including any database in the project "on a whim", without the need of strongly modifying the code.

2.1 Fornasier - Spectra of M Asteroids

This database comes from a survey of metallic asteroids, which acquired spectra of 30 such bodies in the years 2004-2007. Spectra cover the range of wavelengths $[0.4\mu\text{m}, 2.5\mu\text{m}]$ (in some cases only a smaller range of roughly

$[0.4\mu\text{m}, 1\mu\text{m}]$), and the reflectance values are normalized at 0.55μm.

Observations were done with *Italian Telescopio Nazionale Galileo* (TNG) of the European Northern Observatory in La Palma, Spain, and the New Technology Telescope (NTT) of the European Southern Observatory in Chile. Six asteroids were also observed with NASA Infrared Telescope Facility (IRTF) in Hawaii.

The properties of each observing instrument, the time and conditions (airmass) of each observation, are all reported in Fornasier *et al.* 2010. The paper elaborates on data reduction used to obtain composite spectra by putting together different observations. Files related to individual observations (from each telescope) are also available in the database. Although the code developed could download those files in the same manner, accessing composite spectra (joint data from observing with multiple telescopes at the same time) has been deemed sufficient for the purposes of this project, and only data from a relevant link has been downloaded.

In the case of one asteroid, 125 Liberatrix, three spectra were acquired, corresponding to different stages of the asteroid's rotation around it's own axis. The aim was to see if the asteroid's surface had variety dependent on the rotational phase, but the spectra did not differ a lot (see Figure 1).

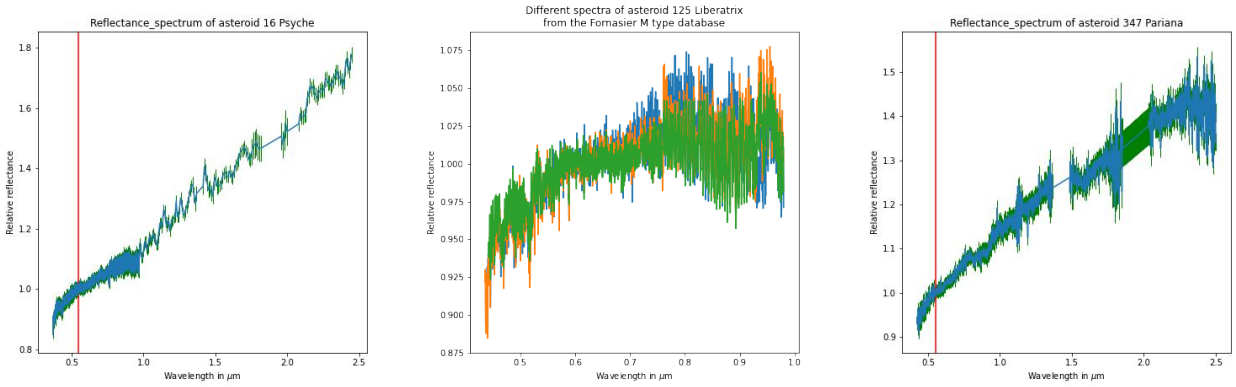

Figure 1: Examples of spectra obtained from the Fornasier M type asteroid database. **Left:** spectrum of 16 Psyche, a target asteroid of a recently launched NASA mission (NASA n.d.(a)). The red vertical line marks here, and on all subsequent plots in this report, the normalization wavelenth assumed in a given database. **Center:** Comparison of three spectra of 125 Liberatrix acquired at different rotational phases. For clarity, error bars and the normalization wavelength are not shown on this plot. Notice that the range of wavelengths covered for this asteroid was smaller than for the other two. Checking Table 1 in Fornasier *et al.* 2010 one can verify that Liberatrix has only been observed with the *Dolores* instrument of the TNG telescope, and not with instruments used for infrared observations in this study. **Right:** 347 Pariana, one of the asteroids for which an absorption band near 0.9μm was identified.

The main spectral features to be found among some of these M-type asteroids are:

- An 0.9μm absorption band attributed to low iron orthopyroxenes. Noticable for example in the case of 347 Pariana on the right of Figure 1
- An 0.43μm absorption band, the origin of which is disputable (theories suggest chlorites, magnesium rich serpentines or pyroxenes such as pigeonite or augite).

2.2 Gartrelle - Spectra of D asteroids

The database Gartrelle *et al.* 2021a features spectra of 25 D-type asteroids from varying Solar System locations. They cover the visible near infrared range ($[0.69\mu\text{m}, 2.5\mu\text{m}]$), with reflectance normalized to the value at 1.5μm. The spectra were obtained from NASA/IRTF on Mauna Kea between 2016-2019.

As explained in Gartrelle *et al.* 2021b, the inspiration for looking into the D-type asteroids had stemmed from the fact that very little is known about them. D-type asteroids are characterized by steep spectral slopes (they appear very red), low albedos and lack of features in the spectrum (no characteristic mineral absorption bands). Moreover, these

objects appear primarily among Trojans (around L4 and L5 Lagrange points of Jupiter, with differences between the two populations associated with different level of stability in the two regions) and in the outer part of the asteroid belt (although there are several NEAs of D-type known as well), with only three confirmed cases of meteorites that are considered to have come from D-type asteroids. Although numerous correlations and conjectures have been put forward, such as D-type asteroids being extinct cometary nuclei, the spectral class remains shrouded in a veil of mystery. In Gartrelle *et al.* 2021b, spectra from Gartrelle *et al.* 2021a were used to look for correlations between the redness and heliocentric distance for D-type asteroids, concluding that the reddening of a spectrum is stronger for objects closer to the Sun.

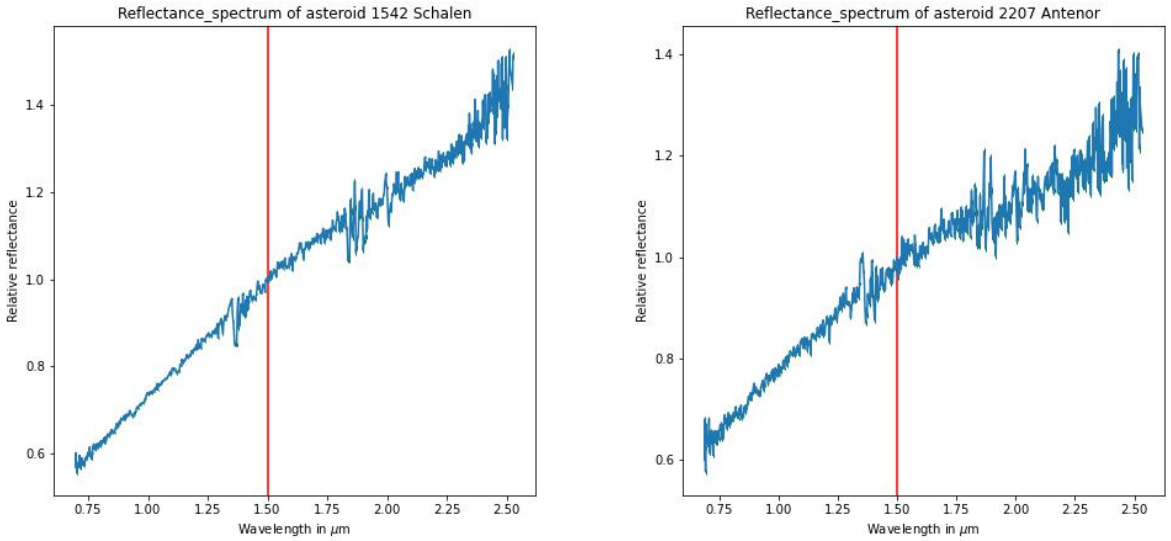

Figure 2: Examples of D-type asteroid spectra from Gartrelle *et al.* 2021a. **Left:** Spectrum of 1542 Schalen from the middle part of the Main Asteroid Belt. The heliocentric distance at the moment of observation was 3.094 AU. **Right:** Spectrum of 2207 Antenor, a Trojan object from the Jupiter L5 point group. The heliocentric distance at the moment of observation was 5.15 AU. Orbits of both objects have low inclinations (under 6°). Notice how the range of ordinates on the left plot extends to higher values than on the right plot, exemplifying an anticorrelation between the redness of the spectrum and the heliocentric distance for D-type asteroids. The difference, while noticeable in this example, is not a very stark one, and advanced statistical methods needed to be applied in Gartrelle *et al.* 2021b to confirm the existence of a trend.

2.3 PRIMitive Asteroids Spectroscopic Survey Library (PRIMASS-L)

This database contains spectra of about 642 asteroids from families and groups that had sparsely been studied before (8 families of inner-belt asteroids and 4 families of outer-belt asteroids). Spectra cover the range of wavelengths $[0.34\mu\text{m}, 0.92\mu\text{m}]$ and are normalized at 0.55μm.

The majority of the visible range spectra come from the Gran Telescopio Canarias (GTC) in La Palma, Spain. Most of the near-infrared spectra come both from the Telescopio Nazionale Galileo (TNG), and the NASA Infrared Telescope Facility. Fewer near-infrared spectra were obtained, therefore for some of the asteroids the wavelength range ends shortly over 0.7μm.

It is important to point out that the PRIMASS survey had started in 2010 and is ongoing, so more and more spectra will be available through this database. In parallel with the survey, a dedicated *Pyton* library for asteroid reflectance spectra analysis, called CANA, had been developed (De Pra *et al.* 2018), which is also used in some aspects of this project.

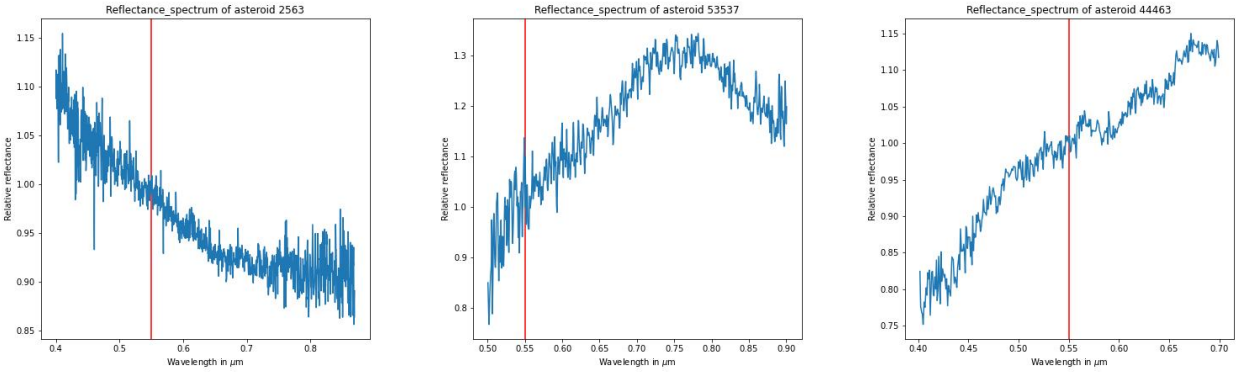

Figure 3: Examples of spectra obtained from the PRIMASS-L database. This database contains a lot of different spectra, exhibiting a large variety of spectral features (instances of every S. J. Bus *et al.* 2008 taxonomic class can be found within the database). The three spectra shown here belong to classes which have a relatively high number of representatives: B-type (**left**), A-type (**center**) and L-type (**right**).

2.4 SMASS and SMASS II

The Small Main-Belt Asteroid Spectroscopic Survey (SMASS) had two stages, which went on for over 10 years, commencing in 1990. Observations were primarily made using the Hiltner telescope, located at the Michigan-Dartmouth-MIT (MDM) Observatory in Arizona. Spectra of 316 asteroids were published in the first phase of the project and spectra of 1341 asteroids in the second phase (these are not disjoint sets, some asteroids appear in both databases, see examples in 10). The wavelength range covered is $[0.42\mu\text{m}, 1.02\mu\text{m}]$ and normalization is performed at 0.55μm.

As the name suggests, the main goal of the survey was to accumulate spectroscopic data for small (< 20 km for a majority of objects observed by SMASS) asteroids in the Main Belt. The main objectives behind that were (Xu *et al.* 1995) increasing fractions of asteroids with known taxonomic classifications (especially for NEAs, see Popescu *et al.* 2019), looking for meteorite analogs and investigating characteristic absorption features, which are known to exhibit more versatility in that regard, hinting at higher compositional diversity.

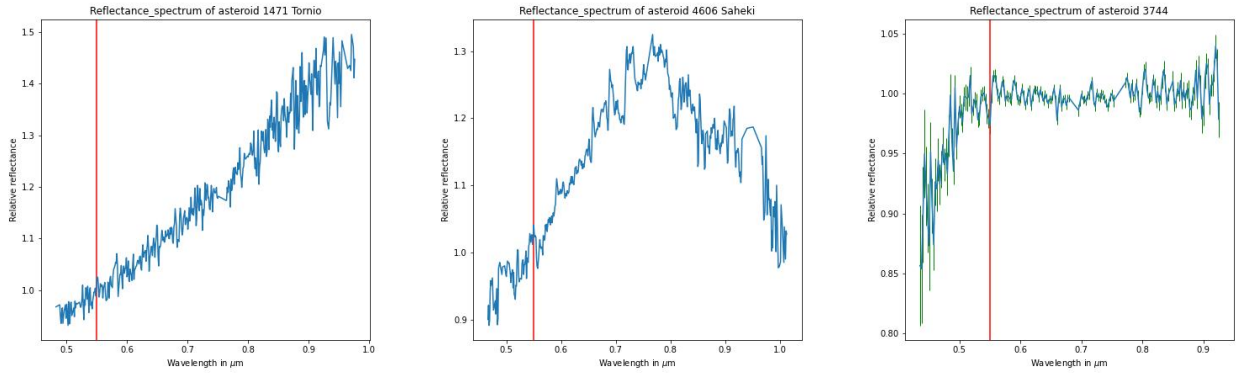

Figure 4: Examples of spectra obtained from the SMASS and SMASS II databases. Just like PRIMASS-L (2.3), they contains data for asteroids of all kinds of spectral types. Representatives of the three commonly encountered classes are shown here: D-type (**left**), A-type (**center**) and K-type (**right**). The plot on the right comes from the database of SMASS II, as we can see this database, unlike the one from SMASS, provides error bars on relative reflectance, but does not contain asteroid names in the data files.

2.5 Near Infrared IRTF spectra

Although the surveys 2.3 - 2.4 have the advantage of large size and variety of available spectra, the usefulness for analysis, especially in the context of mineralogical content, may be somewhat limited due to the limited wavelength ranges covered there - the upper bound at 1.02μ m excludes potential 2μ m pyroxene absorption band. Since the study of pyroxene and olivine content is one of underpinning aspects of asteroid research, it seemed essential to include a larger infrared-oriented dataset, which would cover the range $\sim 2\mu$ m for numerous asteroids of different spectral classes.

The dataset S. J. Bus 2009, comprising low-resolution, near-infrared spectra obtained with the *SpeX* spectrograph at NASA/IRTF on Mauna Kea, turned out to be a perfect candidate. The database covers a total of 173 asteroids, from most of the S. J. Bus *et al.* 2008 taxonomic classes, and the range of wavelengths is $[0.65\mu\text{m}, 2.55\mu\text{m}]$. The data has been acquired by several different research groups, therefore no single focus can be defined, and one needs to be aware that the methods which had been used for calibration for the Earth's atmosphere's influence may differ from one observation to another. Additionally, most of the asteroids have more than one observation, for some of them there are up to 24 distinct (albeit similar) spectra available in the database.

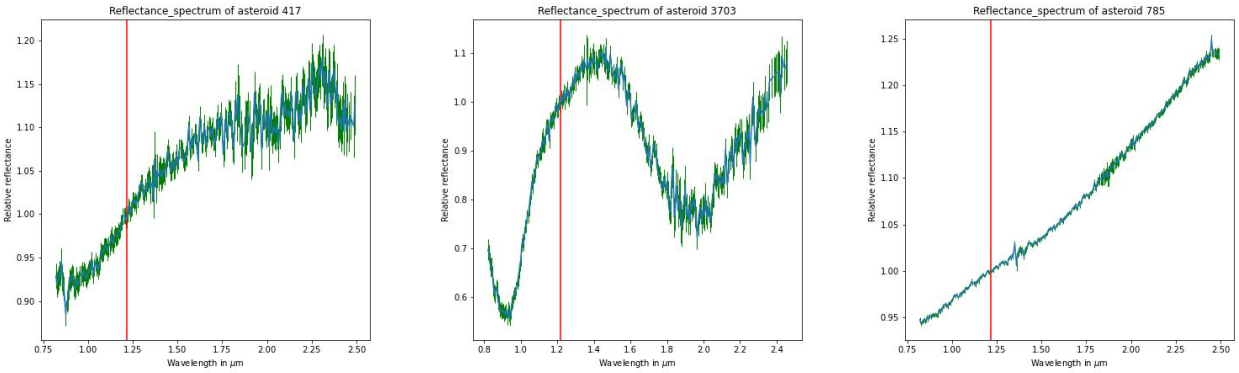

Figure 5: Examples of spectra obtained from the IRFT database. The classes with the most representatives are: D-type (**left**), V-type (**center**) and X-type (**right**). Notice the clear $\sim 2\mu$ m band on the central plot, which would have been missed for a V-type asteroid when looking in the range covered by PRIMASS-L and SMASS.

A huge advantage of the IRTF database is that it provides an observation date for each spectrum in the file name, which allows to access it (see section 3) and use for making accurate phase angle corrections (see section 7) when conducting mineralogical analysis (see section 9.2).

3 Downloading and parsing data

The first step needed to analyze a large dataset is to pre-process it so as to be able to access all the information inside the computer program. The chosen programming language for this project was *Python*, therefore the desire is to store information from data files in *NumPy* arrays. There were multiple challenges related primarily to inconsistencies between ways in which this data is stored in different databases. The full description of the parameters and flags with which the code handles these exceptions, can be found in Morawski 2023-2024. Here, only a conceptual description of the challenges, and ways in which they were addressed, will be provided:

1. Accessing files from a given database:
    - The desired functionality was to be able to provide a single *url* link to a function and be able to directly download all data files with reflectance spectra available within the database, and only those files.
    - Some databases, together with the spectra, contained other auxiliary files. For example, the Gartrelle database (2.2) contains files with observational parameters, as well as an inventory spreadsheet. A consistent pattern is that the files with a spectra always contain the asteroid number in the file name, at a fixed position in the name (usually the beginning). Therefore, it was possible to filter out files which do not contain spectra by checking if a relevant character of the filename is numeric or not.

- In some databases (2.3, second stage of 2.4, 2.5) spectra are not in a single directory, but are subdivided among directories. To be able to identify and access the desired files, a function was created that runs a recursive search among subdirectories to retrieve all relevant links to available spectra. A distinction between the case of databases with/without subdirectories has initially been made, but ultimately a flag differentiating between the two cases has proven redundant and was removed from the code, leading to a higher homogeneity of approach to databases (fewer parameters needed to be adjusted by the user between different cases).

2. Handling different data file types and formats:

- In some cases (2.3) filenames keep an equal length by introducing additional zeroes in front of the asteroid number. To be consistent from one database to another (so that identical asteroid numbers occuring in two different databases would not be considered as different just because one is preceded by a sequence of zeroes), the code disregards initial characters as long as something other than 0 comes up in the filename string. This issue is handled automatically and does not need to be recognized by the user.
- In the SMASS databases (2.4), some of the files have wavelengths in Angstroms instead of microns. As it was the case for all SMASS I files, the initial approach was to handle this by adding a *wavelength_factor* parameter, equal to 1 by default but set to 10^{-4} for SMASS, to multiply all the wavelengths. However, it later turned out that in SMASS II it only applies to a part of the files. The only viable solution seemed to be a brutal approach of checking if the wavelength range starts in thousands, and multiplying by 10^{-4} if that's the case (when downloading each individual spectrum). While initially treated with reluctance, this way of handling the micron/Angstrom inconsistency may ultimately be deemed advantageous, as it lines well with the objective of homogeneous treatment. Indeed, there should be no concern of accidentally misinterpreting the data through this approach - no data file could possibly claim to be showing a reflectance spectrum in microns and have wavelength values of the order of $10^3 - 10^4$. This discrepancy can now by handled by the code, without any need for the user to indicate the need explicitely. One could imagine that some hypothetical database could also use nanometers as a unit, which would then make things even more challenging. In principle, two ways of handling that could be imagined:
  - Define an "expected range" of wavelengths, as for a typical reflectance spectrum we expect it to cover a subrange of visible and near infrared light ($\in [\sim 0.5\mu\text{m}, \sim 2.5\mu\text{m}]$). Then, for each data file the code could look for the most fitting $n \in \mathbb{Z}$, $n = 0$ by default, such that the wavelengths would fall in that range when multiplied by 10^n.
  - Assuming that databases other than SMASS II have a consistent wavelength unit within themselves, the code could still give an option of manually providing a multiplication factor for an entire database, as was initially done for SMASS.

  Since the situation is hypothetical anyways, a simpler case of the latter solution has been adopted for this project. One (also rather hypothetical) advantage would be that any multiplication factor, not only a power of 10, can be used. The former solution would not take too much effort to implement on top of that, should the need ever arise.
- As mentioned in section 2, while in some databases the files contain three columns (wavelengths, reflectance and reflectance errors), in some other ones (2.2,2.3, first stage of 2.4) the errors are not provided. This is the source of a visual discrepancy between Figures 1, 4 right, 5 versus Figures 2, 3, 4 left and center. Although the errors are hardly being used in this project, foresight encourages saving them whenever available, so that each spectrum object would store three *NumPy* arrays corresponding to wavelengths, reflectance and errors. This is handled in the code at a file processing level by checking the number of the columns in a file and setting errors to an array of zeroes if there are only 2 columns.
- Default data file extension is *.tab*, but in some cases it is different, such as *.csv* in 2.2. The data in this format is separated by comas instead of spaces. To handle these situations, a keyword containing the file extension, with a default value set to *.tab*, can be modified to a database-relevant value. This step needs to be done manually by the user, although if a higher degree of homogeneity were required in the future for a big data project, one could imagine automating it by recognizing the most commonly encountered file extension within a directory. The delimiter between columns is determined automatically at the level of processing individual files, based on the first character which is not a number or a dot.
- Some databases (2.1,2.2,first stage of 2.4) provide the name of the asteroid in the filename, while others only provide a number. It felt pleasant to extract the names of the asteroids. To do this whenever possible and save these names in a global dictionary, a relevant flag needs to be set to **True** for a given database.

With the current version of the code, doing so will lead to an error if the file names do not contain asteroid names. However, the opposite is not true - one can still successfully download all the data without trying to extract asteroid names from the file names. Therefore, a hypothetical big data project would not require setting this flag individually for each database. Ultimately, while useful in the earlier stages of the project, the feature of asteroid name extraction has been largely made redundant by an alternative way of accessing auxiliary information about asteroids which will be described in section 4.

- Perhaps most importantly, although discovered quite late in the project, in the case of IRTF database (2.5), observation dates are provided in the filenames and should be extracted for the sake of making accurate temperature and phase angle corrections (see section 7) when performing a mineralogical analysis. A flag allowing extraction of this information has been added to the code, and a relevant attribute of the spectrum object stores this information for every spectrum, with nan instead if the information is not available or was not extracted

3. Evaluating relevant database-specific information that should be saved for each spectrum:

   - One can imagine that for certain research projects, a set of database-level parameters, informing about ways in which the data had been acquired and processed, should be saved together with each spectrum. In this project, one obvious parameter of such kind has been considered - the normalization wavelength. Let us not forget that the data stored in a file corresponds to relative reflectance, as the actual reflectance value depends on the solid angle of observation and is generally not as straightforward to infer from observations (see introduction of Morawski 2023). The spectra are normalized to 1 by dividing irradiance at each wavelength λ by the value at a fixed λ_0 $\left(r_{\text{relative}} = \frac{I_\lambda}{I_{\lambda_0}}\right)$. Although the normalization wavelength λ_0 could often be inferred from the data (for example when looking at the D-type asteroid spectra on Figure 2 one may estimate very well a unique (aside from minor fluctuation in a small range) abssica value for which the ordinate assumes 1), the spectra are not always monotonous and generally it can be beneficial to have the knowledge regarding the wavelength at which the normalization took place.
   - In order to find information about the normalization wavelength which had been used in a given database, one needs to read description text files for each database. This type of research into databases properties is anyways beneficial for learning about the focus of each survey and former use of the data - for the databases studied in this project, a similar research has been a seed of short descriptions provided in subsections of 2.
   - In the current shape of the code, the normalization wavelength is a required parameter which needs to be provided with a database. A hypothetical future big data project would presumably require a modification in one of the following directions:
     (a) Make the normalization wavelength an optional parameter only
     (b) Infer the normalization wavelength automatically by analyzing several files from a database and finding a wavelength at which relative reflectance is always equal to $r_{\text{relative}} = 1$
     (c) Renormalize all spectra at a new, fixed wavelength. This approach may perhaps be the most viable, as it would enforce a higher degree of homogeneity without a loss of scientific accuracy (all it would do is multiplying by another constant). A prominent candidate for such new λ_0 would be 0.55μm, as it is the most commonly used normalization wavelength (2.1,2.3,2.4).

     Additionally, in some rare cases more sophisticated normalization techniques can be found, such as normalizing to an average over a range of wavelengths, which could lead to inconsistent λ_0 value within a single database. Ultimately, including databases with such normalization methodology in the project in it's current stage feels precarious, and a modification of the approach to the way normalization information is stored should perhaps be considered. On the other hand, however, a proxy of a center of the averaging interval could be provided as normalization wavelength in the code for such an example, and generally speaking a proxy could still be helpful for analysing the data.

4. Avoiding repeated queries to the same database:

   - The initial focus on online databases and *url* links in the project has proven to be a double-edged sword. While it allowed achieving high flexibility and automation of downloading data stored in different places and different formats (all the aforementioned challenges), it resulted in initial functions being able to access data from an online database only. Consequently, each time the *Jupyter Notebook* (Morawski 2023-2024) would be opened, the same queries to an online database would be made yet another time, and a bulk of data would be downloaded to the same directory, overwriting data which had already been saved.
   - This lead to multiple inconveniences:

  - dependence on reliable internet network to progress with the project
  - slower execution times, especially for the cases of databases with an intricate network of subdirectories - a recursive search of those, while exhaustive and functional for any structural makeup of the subdirectory network, is slow in terms of computation time
  - wasteful use of bandwidth for accessing the same data over and over
- A solution has been utilized with an introduction of an additional flag *load_previously_downloaded* in the code. When set to **True**, it will disregard *url* links and access data from an existent directory instead. The sole disadvantage is that for the sake of unification, the initial run of the downloading function does not save the names of asteroids in the filenames, but only the numbers, and this information is lost when loading data from there. With advances explained in section 4, even this subtle inconvenience ceases to be of any relevance.
- An analogous problem of losing information could arise for observation dates in the case of IRTF database. To avoid it, functions were created which would save a list of observation dates for each asteroid in a text file during the first download, and could extract it from that file during later runs. In that case, phase angles and heliocentric distances are being extracted with a method described in 7.4 in the first run and stored in the same file. A change of project's internal file naming convention was also necessary, so that the observation date would be saved in the file name when available, in order to be able to link each spectrum with observation-specific information.

Having achieved a desired functionality of loading reflectance spectra data from different databases, it was possible to move on with the project and write functions that would perform different types of data analysis, which will be described in sections 5-10. Beforehand, let's take a quick look at several helper functionalities which played an important role in the project even though they do not bring new scientific value in themselves:

- **Plotting.** A function to generate plots such as those shown on Figures 1-5 was created. It has also served as a foundation for other, more sophisticated plots, such as on Figures 29-31. The plotting function for a spectrum object in the code:
  - plots relative reflectance as a function of wavelength
  - may optionally show the error bars on relative reflectance, if this information is provided in the data file and the user does not suppress plotting them by changing a flag *mark_errors*
  - marks the normalization wavelength with a vertical red line, unless suppressed by the user (changing a flag *mark_normalization_wavelength*)
  - may save the plot as an image on the computer, if the user wishes to do so, otherwise it presents the plot inside the *Jupyter Notebook*
  - instead of creating an independent plot may receive/output a plot which can be combined with other plots
- **Renormalizing.** For uses such as comparisons between different spectra, it is convenient to renormalize them at a shared normalization wavelength λ_0'. Additionally, it may in some cases be convenient to multiply an entire spectrum by an arbitrary non-zero number. The renormalizing function fulfills both of these objectives, and may either create a new spectrum object, or only return a renormalized reflectance array.
- **Generating a CANA *spectrum* object.** For parts of the project which use functions from the CANA library (De Pra *et al.* 2018), it is necessary to create an object compatible with the class used in this library. The conversion is rather straightforward.
- **Trimming the range of wavelengths.** In some cases, it may be advantageous to plot or analyze only a section of the spectrum, corresponding to a subrange of wavelengths. The trimming functions trims the wavelengths, reflectance and error arrays to an arbitrary range provided, and either returns the trimmed arrays, or a new object.

4 Accessing additional information about the asteroids

A striking bottleneck on any hypothetical project dealing with spectra of hundreds or thousands asteroids would be inability of accessing other parameters of these asteroids, which are not spectroscopic but may yet prove to have correlations with spectral features or mineral compositions. Most importantly, orbital parameters, as they carry information about the asteroid's position in the Solar System, which could then by tied with probable composition through

evolutional models, would be such a set of parameters one could desire to extract for each asteroid in a study in an automated way. However, in this project a similar conundrum has already occurred for a much simpler parameter of no scientific value - the asteroid names. While not all of the asteroids have been given names, and it would be virtually impossible to assign them all at the same pace as the discoveries are currently made, many of the asteroids for which we have reflectance spectra also have been named. The desire to associate numbers with names and inconveniences encountered with the initial approach have already been mentioned in section 3.

Orbital, observational and any other relevant parameters of asteroids can be found in the Small Body Database (N. JetPropulsionLaboratory n.d.(a)). However, a simple lookup interface does not allow one to efficiently extract relevant parameters for a larger set of asteroids. After a lot of struggle, an alternative solution has ultimately been found to automize acquisition of information from this database. The code can now do so, for a newly encountered asteroid number, by sending a query of the following type:

https://ssd-api.jpl.nasa.gov/sbdb_query.api?fields=name,a,e,i,diameter,rot_per,albedo&sb-cdata=
*%7B%20%22AND%22%20%3A%20%5B%20%22spkid%7CEQ%7C**20000001**%22%20%5D%20%7D%0A*

where:

- the beginning of the link signalizes a request to the database query API
- *fields=* is followed by the list of keys for parameters which we want to extract from the database:
  - *name* - the name of an asteroid
  - a - the semi-major axis of the orbit
  - e - the eccentricity of the orbit
  - i - the inclination of the orbit in degrees
  - *diameter* - the effective diameter of the asteroid in km (an approximation for a diameter of a sphere of the same volume)
  - *rot_per* - synodic rotational period of the body in hours
  - *albedo* - albedo of the object

  these parameters were chosen to be extracted in this project as they presumably may have some relationship with observational spectroscopic parameters. Should a study be interested in other parameters, their tags should be identified at N. JetPropulsionLaboratory n.d.(b) and added to the query.
- *sb-cdata=* means "search by custom field data"
- a subsequent string must be a translation of a *JSON* query to a *url* language. Luckily, only the part marked in bold, referring to the *spkid* search parameter, needs to be changed. The value given in the example above corresponds to Ceres, generally this SPK-ID appears to be always equal to $20000000 + X$, where X is the designation number of an asteroid

The downside of this approach is that a single query takes a few seconds to process, so running it for many asteroids individually increases the execution time by a big factor. To avoid similar problems as with repeated downloads of files with spectra at the beginning of the project (see Section 3), functions have immediately been introduced which save the information extracted from the Small Body Database into a file and access it from that file the next time the program is opened.

One last challenge was that the string outputted by this query cannot be properly processed (converted into a dictionary and used to access parameters) if some of these parameters are not present in the database. To avoid losing those parameters which can be extracted, a workaround has been found to automatically find and replace all instances of *null* for *"nan"* in the query output. There was also a special case bug in situation when asteroids were named after a person with a multi-component surname, such as 2019 van Albada. When data was accessed from the database, it has been saved with a space in the text file, and when loading it again in the next execution of the code, it would be treated as a delimiter, leading to wrong assignment of values to parameters and some numerical parameters failing to process the string. The bug was discovered when working on the analysis from section 5.4, and a temporary solution of manually modifying the file for those few cases by replacing spaces with underscores has been applied for efficiency (so as not to be compelled to send hundereds of queries to the Small Body Database again). To avoid similar problems

in the future, the code has then been modified to use comas instead of spaces as a delimiter.

An example of auxiliary data extracted in such an automated manner for all bodies covered in the Fornasier database (2.1) is presented in the Table 1.

number	name	a [AU]	eccentricity	inclination [°]	diameter [km]	rotation period [h]	albedo
110	Lydia	2.733	0.0797	5.96	86.09	10.927	0.1808
125	Liberatrix	2.743	0.0807	4.67	48.418	3.968	0.182
129	Antigone	2.867	0.2129	12.27	113.0	4.9572	0.151
132	Aethra	2.613	0.3871	24.97	42.87	5.1684	0.199
135	Hertha	2.429	0.2072	2.3	79.24	8.403	0.1436
16	Psyche	2.924	0.1342	3.1	226.0	4.196	0.1203
161	Athor	2.38	0.1368	9.06	40.992	7.28	0.23
201	Penelope	2.679	0.1791	5.76	85.877	3.7474	0.04
216	Kleopatra	2.793	0.251	13.12	122.0	5.385	0.1164
22	Kalliope	2.911	0.0987	13.7	167.536	4.1483	0.166
224	Oceana	2.645	0.045	5.85	58.236	9.401	0.166
250	Bettina	3.144	0.1363	12.82	120.995	5.0545	0.112
325	Heidelberga	3.216	0.1513	8.57	75.72	6.737	0.1068
338	Budrosa	2.912	0.0181	6.04	50.506	4.6084	0.276
347	Pariana	2.615	0.1638	11.69	48.615	4.0529	0.19
369	Aeria	2.649	0.0974	12.72	73.767	4.778	0.127
382	Dodona	3.122	0.1704	7.39	65.209	4.113	0.129
418	Alemannia	2.593	0.1188	6.82	40.33	4.671	0.201
441	Bathilde	2.806	0.0817	8.16	65.131	10.446	0.204
498	Tokio	2.652	0.2238	9.5	81.83	41.85	0.0694
516	Amherstia	2.677	0.2752	12.95	65.144	7.4842	0.202
55	Pandora	2.759	0.1444	7.18	84.794	4.804	0.204
558	Carmen	2.907	0.0383	8.37	54.811	11.387	0.131
69	Hesperia	2.976	0.17	8.59	138.13	5.655	0.1402
755	Quintilla	3.187	0.1355	3.24	41.21	4.552	0.124
785	Zwetana	2.572	0.2086	12.77	49.46	8.8882	0.12
849	Ara	3.144	0.201	19.54	80.756	4.116	0.186
860	Ursina	2.797	0.1084	13.29	34.561	9.386	0.116
872	Holda	2.732	0.0798	7.39	34.431	5.945	0.165
97	Klotho	2.668	0.2581	11.78	100.717	35.15	0.128

Table 1: Information extracted from the Small Body Database (N. JetPropulsionLaboratory n.d.(a) for asteroids extracted from the Fornasier database (2.1)

A similar approach has been utilized to access observation specific information related to the asteroids position in space when a given spectrum has been taken. Since that step of the project is very tightly linked with corrections to absorption band area ratio, it will be described in the section 7 dedicated to that topic.

5 Taxonomic classification

5.1 Theoretical background

One forthright practice which can be administered to a large dataset, is classifying members of that set into subcategories and checking how many members belong to each of them. For the case of reflectance spectra of asteroids, the taxonomic classification scheme of S. J. Bus *et al.* 2008 offers a ready system to categorize asteroid spectra. It introduces the following structure of categories and subcategories (see Schelte J. Bus and Richard P. Binzel 2002):

- **C-complex** - commonly associated with carbonaceous chondrite meteorites, asteroids from the C-complex have low albedoes, and their spectra have low slopes, with some absorption features, which, however, are rather faint, with a drop in relative reflectance of only a few percent in the center of a band. The most notable of those

features is a band at 0.7μm, indicating the presence of phyllosilicates likely due to aqueous alteration (Vilas and Gaffey 1989). The complex is subdivided into following classes:

  - **B** - remarkable as the only spectral class with a tendency for the reflectance to be a decreasing function of wavelength
  - **C** - generally flat and featureless except for weak absorption band below 0.55μm
  - **Cb** - very flat and featureless
  - **Cg** - with a strong absorption below 0.55μm and occasionally another weaker feature around 0.85μm
  - **Cgh** - similar to Cg but with an addition of a moderate absorption band centered near 0.7μm
  - **Ch** - similar to C but with a smiliar addition of a moderate feature near 0.7μm

- **S-complex** - asteroids belonging to S-complex are expected to have a silicaceous composition (Chapman, Morrison, and Zellner 1975). They are characterized by spectra with moderate olivine and pyroxene absorption features around 1μm and 2μm. The following subclasses are delineated based on the strength of a 0.9μm drop, indicative of the 1μm band:

  - **S** - with a broad, but shallow absorption feature centered near 0.63μm
  - **Sa** - intermediate between S and class A from end members (see below)
  - **Sq**, **Sr**, **Sv** - analogously as Sa, bridging between S and end member classes of Q, R and V respectively

- **X-complex** - a class of asteroids with moderately sloped spectra with few features or featureless. It is known to be compositionally degenerate as it comprises both very dark (low-albedo) and bright (albedo around 0.5) asteroids. The following subclasses are distinguished:

  - **X** - generally featureless and moderate, with an occasional shallow absorption feature longward of 0.85μm
  - **Xc** - mostly featureless
  - **Xk** - similar to Xc, but with a higher slope (redder)
  - **Xe** - with a remarkable absorption feature around 0.49μm which may be associated with troilite (FeS)

- **End members** - this broad range comprises asteroid with spectra which exhibit more extreme or distinct spectral features and hence deviate from the other 3 complexes:

  - **T** - mostly featureless, moderately steep shortward of 0.75μm and gradually flattening above
  - **D** - very high spectral slopes, associated with opaques, a lot of question marks on their properties (see 2.2)
  - **Q** - mostly very low iron ordinary chondrites (LL OCs), spectra have a deep, rounded 1μm absorption band
  - **O** - strong, absorption features, especially a feature longward of 0.75μm; contain pyroxenes and olivines
  - **R** - also large absorption feature longward of 0.75μm associated with pyroxenes and olivines, but steeper in the ultraviolet (UV) part of the spectrum then O or S-complex; an additional small absorption feature centered near 0.52μm.
  - **V** - similar to R but with even deeper absorption features, associated with a greater variety of minerals (e. g. plagioclase feldspar)
  - **A** - extremely steep below 0.75μm with a moderate absorption feature above and a subtle feature at 0.63μm, with links to several groups of chondrite and achondrite meteorites
  - **K** - moderately steep below 0.75μm and flat above, with a subtle olivine absorption band
  - **L** - somewhat similar to K but even more featureless, associated with spinel-rich materials

As emphasized in DeMeo *et al.* 2015b, the assignment of an asteroid to a given taxonomic class should be considered as a sort of catalogue labelling rather than a definite statement about the properties of the object. The advantage of taxonomic classification over a detailed mineralogical analysis is that it can be quite reliably performed with a spectrum of lower quality, and hence far more asteroids have been assigned a taxonomic class than analyzed mineralogically in great detail. However, such assignment may not always be definite, and could be contradicted by a future observation, either as a result of observing with higher spectral resolution, extending the spectrum to a broader range of wavelengths, or the fact that an asteroid would be observed at a different rotational phase and a surface irregularity influenced a spectrum.

5.2 Automised classification with CANA

The Bus-DeMeo taxonomy has been implemented in the CANA library (De Pra *et al.* 2018), where a set of exemplary spectra, one for each class, is used for comparison. Each spectrum can be then compared with an exemplary spectrum based on the following χ^2 metric:

$$\chi^2(r, r_c) = \frac{\sum_{i=1}^{n} \left(r_c(\lambda_i) - \frac{r(\lambda_i)\beta(r,r_c)}{r_c(\lambda_i)} \right)^2}{n}$$

where $\lambda_1, \lambda_2, ...\lambda_n$ are the wavelengths at which data on the spectrum, relative reflectance $r(\lambda)$, is available, c is the taxonomic class, r_c is a comparison exemplary spectrum for that class and:

$$\beta(r, r_c) = \frac{\sum_{i=1}^{n} r_c(\lambda_i) r(\lambda_i)}{\sum_{i=1}^{n} r(\lambda_i)^2}$$

In particular, if $\forall_{i \in \{1,2,...,n\}} r_c(\lambda_i) = r(\lambda_i)$, then $\beta = 1$ and $\chi^2(r, r_c) = \frac{\sum_{i=1}^{n} (r_c(\lambda_i)-1)^2}{n}$, which is expected to be the minimum value. Notice that $r_c(\lambda_i) - 1$ are small numbers because r_c is a model spectrum, normalized to 1 somewhere in mid-range. The more different r and r_c are from each other, the more would χ^2 deviate from this minimal value. It does not matter if r is normalized at the same wavelength as the model spectrum r_c or not. Indeed, let $\alpha = \frac{1}{r(\lambda_0)}$, where λ_0 is the wavelength at which the model spectrum r_c has been normalized ($r_c(\lambda_0) = 1$). If r' where the spectrum r normalized at λ_0, then $\forall_{i \in \{1,2,...,n\}} r'(\lambda_i) = \alpha r(\lambda_i)$, and consequently:

$$\beta(r', r_c) = \frac{\sum_{i=1}^{n} r_c(\lambda_i) r'(\lambda_i)}{\sum_{i=1}^{n} r'(\lambda_i)^2} = \frac{\alpha \sum_{i=1}^{n} r_c(\lambda_i) r(\lambda_i)}{\alpha^2 \sum_{i=1}^{n} r(\lambda_i)^2} = \frac{1}{\alpha} \beta(r, r_c)$$

$$\chi^2(r', r_c) = \frac{\sum_{i=1}^{n} \left(r_c(\lambda_i) - \frac{r'(\lambda_i)\beta(r',r_c)}{r_c(\lambda_i)} \right)^2}{n} = \frac{\sum_{i=1}^{n} \left(r_c(\lambda_i) - \frac{\alpha r(\lambda_i) \frac{1}{\alpha}\beta(r,r_c)}{r_c(\lambda_i)} \right)^2}{n} = \frac{\sum_{i=1}^{n} \left(r_c(\lambda_i) - \frac{r(\lambda_i)\beta(r,r_c)}{r_c(\lambda_i)} \right)^2}{n} = \chi^2(r, r_c)$$

so the normalization of the spectrum r does not have any effect of χ^2. To be able to compute χ^2, it is necessary to be able to sample the model spectrum r_c at the same wavelengths as the spectrum r has been measured at. This is achieved in CANA by searching for the nearest matching wavelength between the two arrays.

Based on this metric, running a relevant CANA function on a spectrum object leads to a set of three suggestions of taxonomic classes with lowest values of $\chi^2(c)$. A classification of a single spectrum is then straightforward, by choosing the first item from the suggested list, i. e. the class with the lowest χ^2. However, as it turned out, in numerous cases if an asteroid has multiple observations within a database, the classification does not turn out consistent from one case to the other. Two ways to handle such exceptions were adopted, corresponding to two sides on Figure 6:

1. **Labeling asteroids with inconsistent lowest χ^2 classification as a separate category "Ambiguous".** In this approach, whenever two different spectra of the same asteroid turn out to yield contradicting taxonomic classifications, it is considered to be a failed classification - either due to low quality of the data, different ways the spectra had been processed and corrected for observational effects, or other effects, a sturdy association of the asteroid with a specific taxonomic class is not possible. All such asteroids become uncategorized, or in other words form their own category of asteroids with ambiguous taxonomy.

2. **Trying to predict the most probable taxonomic class based on all the information outputted by CANA.** Even in the cases where taxonomic classification is not fully consistent between different spectra of the same asteroid, the candidate classes generally cover a small set of possibilities, such as two or three different classes that are not extremely different to each other intrinsically (e. g. Sa and A, or T, L and K). If we consider taxonomy as nothing more than a tool for general classification and orientation about spectral properties, it could be beneficial to have a label for each of the asteroids under study, even if some of those may not be optimally chosen. For this, a solution has been implemented to put forward a candidate taxonomic label which would be, in some way, the most probable based on all the files with spectra available. Specifically, all three candidates and their χ^2 values for each spectrum are saved for every spectrum while analyzing the data. After an entire

database has been parsed, a weighted average $\overline{\chi^2}(r_c)$ is calculated for each class c which appeared at least once as:

$$\overline{\chi^2}(r_c) = \frac{\sum_{k=1}^{m_c} w_{j_k} \chi^2(r_c, r_{j_k})}{\sum_{k=1}^{m_c} w_{j_k}}$$

where $m_c \leqslant m$ is the number of spectra for which a given taxonomic class c has been suggested as one of the possibilities (m being the number of all spectra of that asteroid), w_j are weights and $j_k \in \{1, 2, ..., m\}$ are indices of those spectra which had c as one of candidate classes. Several ways of calculating were considered, and their performance evaluated on the example of 125 Liberatrix (see Table 2):

(a) Setting $\forall_{j\in\{1,2,...,m\}} w_j = n_j$, where n_j is the number of datapoints (wavelengths) in the spectrum r_j. Weighing with n_j is necessary to be able to compare between χ^2 for different spectra, because χ^2 has a default factor of $\frac{1}{n}$ and may thus be larger for one spectrum then the other due to the smaller wavelength resolution and not the actual similarity of the spectra. This approach seemed reasonable but it lead to a paradox in the test case of Liberatrix, where a noisier spectrum lead to exclusion of a class which seemed most probable based on the other two, even though it did not even have the ultimately winning class as one of the top three candidates for itself. Admittedly, the weights should perhaps depend on a lower (fractional) power of n, since the numerator of χ^2 includes a sum of n elements and would also be an increasing function of n. While it could be worth investigating what would be the optimal choice of such a fractional power, or if pehaps the weights should in fact be all equal, it would actually not change the paradoxical outcome of the Liberatrix classification (for indeed the weights n_j are already close to each other in this case, see Table 2). The two consequent solutions have been attempted as an improvement method.

(b) Setting $\forall_{j\in\{1,2,...,m_c\}} w_j = \frac{n_j}{\min\limits_{\gamma\in C_j} \chi^2(r_\gamma, r_j)}$, where n_j is the same as above and C_j is a set of three candidate classes with the lowest χ^2 for spectrum j. Such a choice of weights, while remaining faithful to the need to account for $\frac{1}{n}$ scaling mentioned above, also penalizes noisier spectra with lower weights compared to the less noisy ones, since a noisy spectrum will have a higher minimal χ^2. As the case of Liberatrix reveals, this modification was not sufficient to change the outcome. One way to address this would be to raise the minimal χ^2 in the denominator of the weights to a higher power, for example raising to the power of 1.2 or even squaring it. However, the choice of any exponent other than 1 feels extremely arbitrary and forced without a deeper research into that matter alone. While the idea may be fruitful for some applications, in this project a decision was made to refrain from tampering with the weights further. Instead, the following solution has been devised:

(c) Keeping weights from (a), but enforcing $m_c \equiv m$ by setting

$$\forall_{c\in C}\forall_{j\in\{1,...,m\}\setminus\{j_1,...,j_{m_c}\}} \chi^2(r_c, r_j) = \max_{\gamma\in C_j} \chi^2(r_\gamma, r_j) + A \cdot (\max_{\gamma\in C_j} \chi^2(r_\gamma, r_j) - \min_{\gamma\in C_j} \chi^2(r_\gamma, r_j))$$

where $C = \bigcup\limits_{l\in\mathbb{N}, l\leqslant m} C_l$ is the set of taxonomies suggested for a given asteroid as candidates based on at least one spectrum and $A \geqslant 0$. In other words, this methodology sets χ^2 for taxonomies which were not put forward as one of the top three classifications for a given spectrum to be equal to the maximum of χ^2 of the three candidates plus some positive value of the order of discrepancies between χ^2 for other classes. This should be a good way to reflect the fact that these classes for that spectrum surely had higher χ^2, since $\max\limits_{\gamma\in C_j} \chi^2(r_\gamma, r_j)$ is a viable lower bound, and a difference between χ^2 for candidates with higer χ^2 than the lowest three and the maximum of the lowest three should be of the order of $\max\limits_{\gamma\in C_j} \chi^2(r_\gamma, r_j) - \min\limits_{\gamma\in C_j} \chi^2(r_\gamma, r_j)$.

Two choices of A were considered in this project:

i. $A = 0$, corresponding to underestimating all χ^2 extrapolated in this manner

ii. $A = 1$, which would probably be a closer estimation in most cases, but may overstimate in some

Although the second choice is currently opted for in the code, the difference between both treatments is in fact miniscule and and in particular the result for 125 Liberatrix is the same (Table 2). Indeed, the value of A does not have any influence on χ^2 for those classes, which have been present as label candidates for every spectrum. This approach only serves to penalize classes which did not make it to the top 3 for some spectra, with higher A implying stronger penalization.

(d) Ultimately, decision has been made to combine (c) with $A = 1$ and weights from (b) since they each serve their own purpose:

- the former penalizes classes which deviate strongly for some of the measured spectra, even if they match well for others
- the latter penalizes noisier spectra to have lower impact on the weighted average

The final taxonomic label assigned in this scenario to each asteroid is the class c with the lowest value of $\overline{\chi^2}(r_c)$ calculated with the chosen method. Currently, the choice of (d) is adopted in the code. Should a higher flexibility in that regard be desired, it would be possible to modify the code by adding several parameters that would allow to choose among all the methods and their subtle possible modifications mentioned above (in the most general case we could have a similar scenario as (d) but with arbitrary A and weights in the form $w_j = \frac{n_j^a}{\min\limits_{\gamma \in C_j} \chi^2(r_\gamma, r_j)^b}$, where $a, b \in \mathbb{R}^+$ would also be parameters).

Spectrum number (j)	$\chi^2(r, r_X)$	$\chi^2(r, r_{Xc})$	$\chi^2(r, r_{Xe})$	$\chi^2(r, r_{Xk})$	n_j	$\frac{n_j}{\min\limits_{\gamma \in C_j} \chi^2(r_\gamma, r_j)}$
1	$6.361 \cdot 10^{-4}$	$5.695 \cdot 10^{-4}$	-	$5.215 \cdot 10^{-4}$	*2312*	$\mathbf{4.43 \cdot 10^6}$
2	$5.884 \cdot 10^{-4}$	$5.926 \cdot 10^{-4}$	-	$5.588 \cdot 10^{-4}$	**2351**	$4.21 \cdot 10^6$
3	-	$7.320 \cdot 10^{-4}$	$7.661 \cdot 10^{-4}$	$8.170 \cdot 10^{-4}$	2317	$\mathit{3.17 \cdot 10^6}$
Calculation method	$\overline{\chi^2(r_X)}$	$\overline{\chi^2(r_{Xc})}$	$\overline{\chi^2(r_{Xe})}$	$\overline{\chi^2(r_{Xk})}$		
(a)	$\mathbf{6.12 \cdot 10^{-4}}$	$6.31 \cdot 10^{-4}$	$7.66 \cdot 10^{-4}$	$6.32 \cdot 10^{-4}$		
(b)	$\mathbf{6.13 \cdot 10^{-4}}$	$6.21 \cdot 10^{-4}$	$7.66 \cdot 10^{-4}$	$\mathit{6.14 \cdot 10^{-4}}$		
(c), $A = 0$	$6.80 \cdot 10^{-4}$	$\mathbf{6.31 \cdot 10^{-4}}$	$6.65 \cdot 10^{-4}$	$\mathit{6.32 \cdot 10^{-4}}$		
(c), $A = 1$	$7.08 \cdot 10^{-4}$	$\mathbf{6.31 \cdot 10^{-4}}$	$7.14 \cdot 10^{-4}$	$\mathit{6.32 \cdot 10^{-4}}$		
(d)	$6.90 \cdot 10^{-4}$	$6.21 \cdot 10^{-4}$	$7.11 \cdot 10^{-4}$	$\mathbf{6.14 \cdot 10^{-4}}$		

Table 2: **Upper part:** χ^2 values for all candidate taxonomic classes for spectra of Liberatrix 125 from the Fornasier database, as calculated with the function from the CANA library (De Pra *et al.* 2018). Each spectrum receives 3 possible classifications with the lowest χ^2 for that spectra, therefore values are only provided in 3 colums for each case. It makes sense that category Xe is an outlier of sorts, with only one noisy spectrum hinting at that possibility as the second most probable - indeed, spectra of asteroids from the Xe class deviate more noticeably from other X-complex asteroid spectra, and they even have their unique absorption feature (section 5.1), so they should be less likely to be mistaken for when analysing a spectrum of another X-complex category. The last two columns show values of weights adopted in this case when following the scheme in (a)/(c) and (b)/(d), respectively. The relashioship between the weights in each column is emphasized by writing the biggest one in bold and the lowest one in italic. Notice how severely spectrum 3 becomes penalized with the second choice of weights.**Lower part:** averaged $\overline{\chi^2}$ calculated with different methods described in section 5.2. Bold font indicates the lowest value of a given row. Bold italic is used to indicate values which come very close to the lowest one (for Xk in methods (b) and (c)). We can see that method (a) assigns class X, which feels paradoxical since it would not be considered a candidate at all based on spectra 1 and 2 only (Xk would have the lowest averaged $\overline{\chi^2}$ based on the two), nor based on spectrum 3 alone (not even in the top 3 labels for this one) and yet when we put 1, 2 and 3 together we get the lowest $\overline{\chi^2}$ for X just because spectrum 3 is noisier. Method (b) penalizes the noisy spectra, but it does not suffice to change the verdict - $\overline{\chi^2}$, for Xk is now almost as low as for X, but still slightly higher. Methods (c) and (d) successfully exclude the suspicious verdict of class X by penalizing classes X and Xe for not making it to the top 3 suggestions for some of the spectra. Notice that the choice of A does not impact the values of $\overline{\chi^2}$ for classes Xc and Xk, which belonged to top 3 suggestions for all spectra - in particular they are the same in (c) as in (a) and in (d) as in (b). Method (c) slightly favors category Xc, whereas method (d) strongly states that Xk should be the answer. Intuitively, it seems reasonable to choose one of the two but favor Xk, since they both come to top 3 for all spectra, but Xk has lower χ^2 for the less noisy spectra.

5.3 Application to the data

This procedure has been applied to all databases introduced in section 2, leading to Figures 6-10. The histograms are generated in such a way to reflect the order of taxonomic classes in which they were described in this section. Additional empty space is added between classes belonging to different complexes (C, S, X, End members), and a bigger space separates all classified cases from the "ambiguous" category, if the first of the two strategies mentioned above is applied. For the first two databases (2.1,2.2), since only a handful of taxonomic classes is present, the absent ones are not shown on the plot for the sake of clarity. The other ones (2.3,2.4,2.5), coming from larger surveys, include representatives of all or most of the classes, so a space for each one is reserved on the plots in Figure 8-10, even if some

may contain 0 asteroids (Xc, Xe, Q, O and K on Figure 10). Let's take a look at each of those separately.

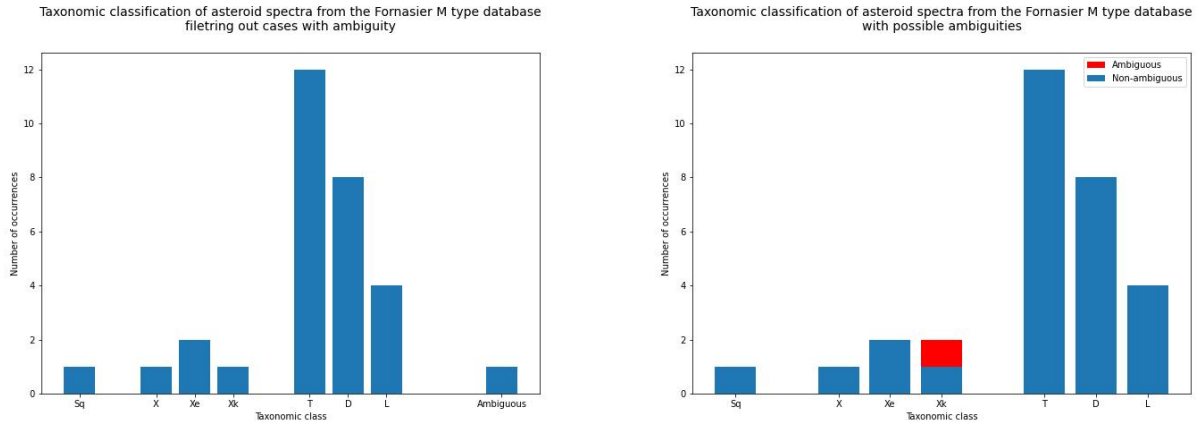

Figure 6: Taxonomic classification of the asteroids which were observed by the survey from 2.1, obtained with two different approaches described in section 5.2. **Left:** method 1, where all asteroids with ambiguities are shown on the side as a separate category. **Right:** method 2(d), which resulted in classifying one ambiguous case (125 Liberatrix) as Xk. In this case, the addition to the bar of the chosen category is included in the plot, but marked with a red color so as to remember that the classification is less concordant than for other asteroids. Bear in mind that only one ambiguous case was possible in Fornasier, because most asteroids had only one observation, with a sole exception of 125 Liberatrix, which had intentionally been observed at different rotational phases. On Figure 1 center we can see that these spectra had a lot of noise, which surely enhanced the likelihood of ambiguous taxonomic classification.

For the Fornasier (2.1) database, a majority of the asteroids are classified into end member classes T, D and L. These classes all have a characteristic of being relatively steep and featureless, which is in accordance with a typical look of the spectrum from the Fornasier database (see Figure 1). A smaller fraction of the cases has been classified somewhere in the X-complex, which also belong to the realm of fairly featureless spectra. There was exactly one case of ambiguous classification, and indeed, more than one would not have been possible since 125 Liberatrix was the only object for which multiple spectra were measured in that study. The second approach lead to assigning different subclasses from the X-complex, depending on the exect choice among the methods proposed in 5.2. A closer look at each spectrum and associated candidate classes with their χ^2 values reveals that only classes from the X-complex were considered, so in the very least we can be confident about the fact that 125 Liberatrix belongs to the X-complex. However, method (a) lead to an absurd situation in which class X is chosen based on the 3 spectra, even though it would not be considered at all when looking at the first two, or the last one only. This very paradox inspired diving into the details of the methodology assumed and refining it until the result matched the intuition and the modifications seemed justifiable (penalizing things they should). The process of refining the method is described in section 5.2 and illustrated by numerical evaluation on the case of 125 Liberatrix in Table 2.

Classification of asteroids from the Gartrelle database, shown on Figure 7, reveals another surprising truth about automated taxonomic classification. Based on both the focus of the survey (see 2.2), as well as a visual inspection of the spectra (e. g. Figure 2, but all plots look quite similar), one would expect that all spectra from this database should be classified as D. And yet, we are faced with two categories, and in fact there are more cases of asteroids classified as A than as D. Apparently, class A is very likely to be a false match and could "pretend" to be other classes. This is also very strikingly visible for SMASS and SMASS II (especially the latter) on Figure 9. Clearly the automatic classification would imply that the majority of asteroids found in SMASS and SMASS II were of class A, which seems statistically improbable, as it is not considered to be a prevalent class and no deliberate focus on this taxonomic class had been put in the survey. Here, we should also keep in mind that the range of wavelengths covered by the SMASS survey is rather limited ($[0.42\mu\text{m}, 1.02\mu\text{m}]$), which makes the classification less reliable for this case.

Histogram of the same kind as Figure 6 right for the first of the larger surveys considered, Primass, is shown on Figure 8. S-complex asteroids are scarce among the Primass dataset, but the other types are distributed quite homogeneously, with most representatives among B, D (which, however, has the highest number of ambiguous cases) and X.

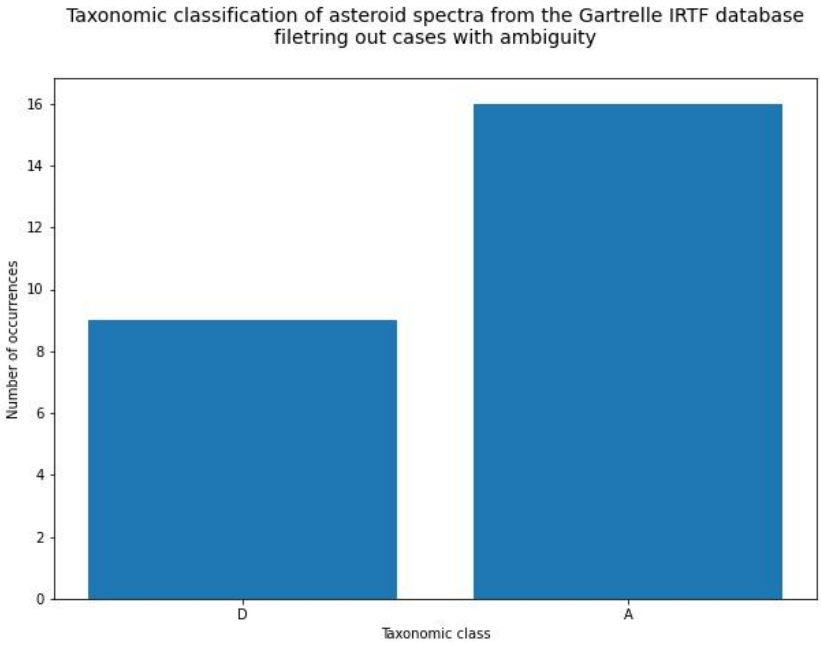

Figure 7: Taxonomic classification histogram for asteroids included in the survey from 2.2, as outputted with methodology described in 5.2. The plot is the same regardless of the ambiguity treatment assumed because in fact there are no asteroids with multiple observations in that database. We should expect them to all belong to the same class D, and yet a majority is misinterpreted to be representatives as A. This could imply that the spectra had been categorized wrongly in the first place, but a more likely explanation is that the automatic classification fails and class A is prone to become a false match.

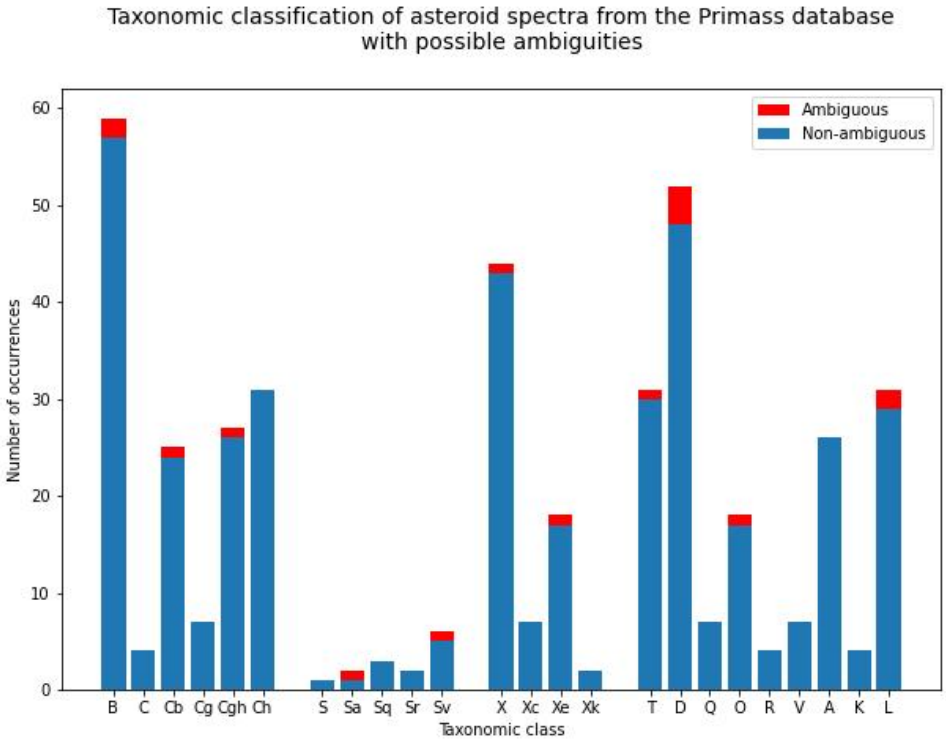

Figure 8: Taxonomic classification histogram for asteroids included in the survey from 2.3, as outputted with methodology described in 5.2. Most asteroids which had multiple spectra available ended up having an ambiguous classification.

SMASS, in it's first phase, seemed to favor end members noticeably, but the distribution becomes closer to uniform with the extention of SMASS II - the histograms are shown on Figure 9, notice that the vertical axis scale is different for SMASS and SMASS II. Few ambiguous cases are assigned to class B for SMASS and K for SMASS II. However, this is not due to high consistency among the data, but due to the fact that very few asteroids had multiple observations in that surveys. This serves as a paramount reminder of the following fact: many more asteroids could end up being classified in an ambiguous manner if repeated observations were made and multiple spectra were available. Indeed, printing out information about the results of the analysis of PRIMASS database spectra by the code revealed, that there was only one case of asteroid with multiple observations available and consistent classification (nameless main belt asteroid 135384, for which 2 spectra were present in the database and they both unanimously lead to labeling this asteroid as X class asteroid), whereas all the other objects with multiple spectra available ended up having ambiguous classification. The situation is somewhat better for IRTF (probably thanks to a larger wavelength range, facilitating accurate taxonomic label assignment), where there are 4 such cases of main belt asteroids with 2 spectra leading to the same conclusion (nameless 135384 assigned to class X, 5771 Somerville assigned to class T, and nameless 107861 and 123979 assigned to class A). Nevertheless, we could forecast that probably if all of the asteroids had multiple observations, most of the classifications would be ambiguous. Perhaps it could help explain the suspiciously high number of A class members in datasets 2.2 and 2.4. In any case, to be able to continue drawing meaningful conclusions from the taxonomic classification, some modifications of approach may be necessary. One idea would be to check if the set of highly probable taxonomic classes for a given asteroids contains only classes which are naturally similar to each other (e. g. X and Xk, Sa and A, L and K etc.), and if so, deem the assignment with method 2(d) from 5.2 as nonambiguous, so that only asteroids which have severely different spectra in different observations would be sieved out as uncategorized. As of now, such a modification is not urgent and beyond the scope of the project.

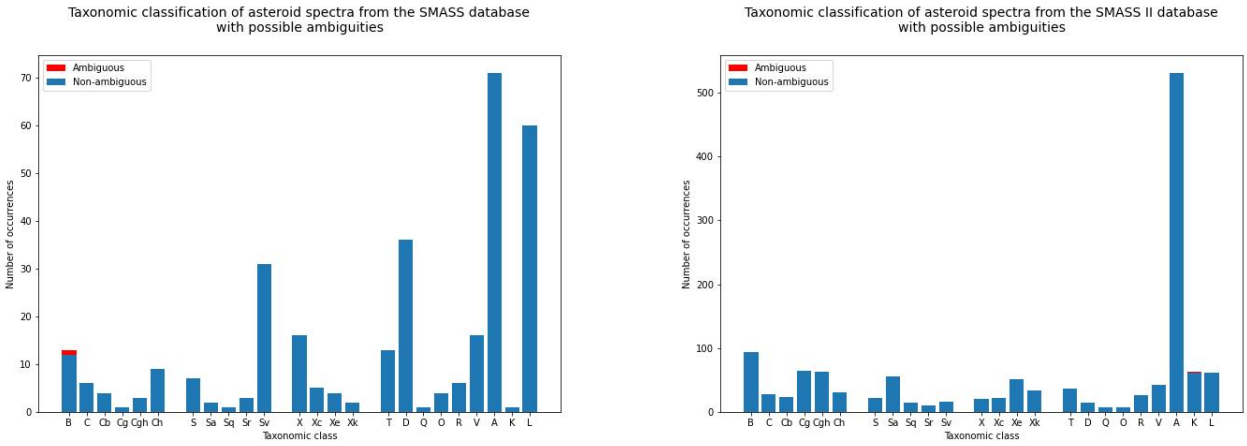

Figure 9: Taxonomic classification histogram for asteroids included in each phase of the survey from 2.4, as outputted with methodology described in 5.2. Notice the suspiciously high number of A class representatives compared to other classes. The scarcity of ambiguous cases is due to the fact that few asteroids had multiple observations in this survey, and not due to excellent matches between different spectra.

The last database IRTF, summarized on Figure 10, does not have a uniform distribution of representatives of taxonomic classes, favoring strongly end member classes D and V. Since V is a very characteristic class with huge absorption bands, and it is rather D that gets mistaken for A than the other way around (and a closer look revealed that most of the asteroids with ambiguity had X as an alternative classification, and only one had A), we can suspect that the assignment is more or less correct and reflects the interest on particular types of spectra expressed more strongly then others in the survey. As mentioned in section 2.5, that dataset is not a result of one broad survey, but it is a mixture of data acquired with the same intrument by different research groups and with various goals in mind. It makes sense that some classes would end up being studied with more detail than others. In particular, the peak in D can be partially due to Gartrelle. Although no overlap is present between datasets 2.2 and 2.5, Gartrelle *et al.* 2021b mentions using another dataset acquired with IRTF, which is presumably a subset of the latter.

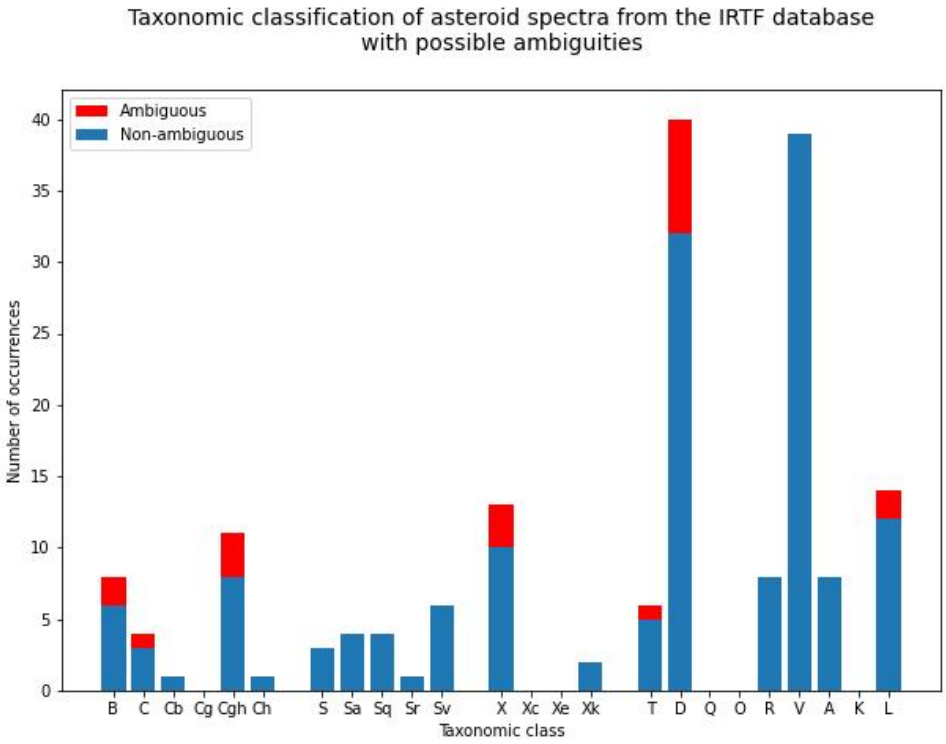

Figure 10: Taxonomic classification histogram for asteroids included in each phase of the survey from 2.5, as outputted with methodology described in 5.2. Notice the focus on classes D and V in that dataset.

5.4 Taxonomic classification in the space of orbital and physical parameters

It is definitely worthwhile to venture a look into how taxonomic classes relate to other (non-spectroscopic) parameters of the asteroids. Having a ready scheme to obtain the latter, as described in section 4, it was easy to add the functionality of creating a scatter plot of asteroids with their taxonomic classes marked. For this part o the project, ambiguities discussed in previous sections were not diferentiated, since it would lead to overly complicated plots without adding much scientiic value, we have learned from the analysis in section 5.3 that anyways majority of classifications were not ambiguous, if only due to only a single observation being available for most asteroids in each database. It was then a matter of running a loop over all databases and pairs of parameters of interest to generate plots presenting the classification of each database in the parameter space defined by each pair. Here, only a handful of most interesting cases will be shown and discussed. In particular, databases of Fornasier (2.1) and Gartrelle (2.2) are of no interest in this regard because they contain too few asteroids to see meaningful trends or clusters on a scatter plot. Additionally, the IRTF dataset can be considered as the one with most reliably assigned taxonomic labels, since it covers a much bigger range of wavelengths (up to 2.55μm, whereas Primass and SMASS end around 1μm). As we are about to find out, it also exhibits the most interesting feature of a V class object cluster in the $a - d$ parameter space. Therefore, this section will focus on exploring that database.

Figure 11 presents taxonomic classification of the IRTF dataset (2.5), juxtaposing semi-major axis against the diameter and eccentricity. The following observations can be inferred from analysing the plots:

- **General properties.** Most of the bodies included in the IRTF dataset are main belt asteroids with a size under 250 km and orbital eccentricities varying more or less uniformly in the $\sim [0, 0.3]$ range.
- **V class cluster.** Taxonomic class V asteroids present in this dataset, except for one outlier (Vesta), all form a very clearly defined cluster of asteroids with very small sizes, and semi-major axes constrained to a $\sim [2.2\text{AU}, 2.5\text{AU}]$ range (the largest of those turns out to be 956 Elisa with a diameter of 10.474km and semi-major axis 2.297 AU, and the smallest 27343 Deannashea with a diameter of 2.864km). No particular dependence on eccentricity can be observed. A look on the semi-major axis - diameter parameter space plot hints at a slightly increasing linear trend in that relationship. A close-up look on that region, with the best linear fit, is presented

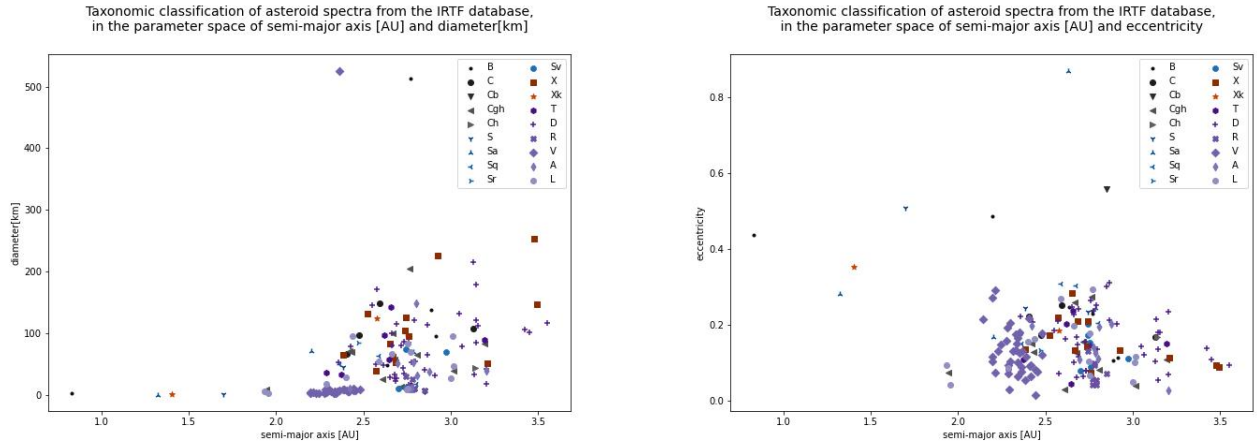

Figure 11: The taxonomic classifications of objects from the IRTF (2.5) database in the parameter space of semi-major axis juxtaposed against diameter (**left**) and eccentricity (**right**). Objects belonging to the same larger category (C-complex, S-complex, X-complex, end members) are shown in shared hue of varying saturation, but with pairwise distinct ,marker shapes. Bear in mind that some of the asteroids present in the database are omitted on the left plot because no information about their size is available. Several prominent outliers and a very clear cluster of small V class objects can be identified, as discussed in detail in section 5.4.

on Figure 12. We can see that the perceived linear trend is mostly due to the fact that the asteroids of this class with the smallest semi-major axes all have small diameters (3703 Volkonskaya, 2579 Spartacus 9553 Colas), while the bodies of the largest semi-major axes have big diameters (3703 Volkonskaya, 4215 Kamo, 2566 Kirghizia). However, the central part o the plot shows a huge spread, in fact no trend whatsoever could be seen on Figure 12 if we were to limit it to the range of, say, $[2.275\text{AU}, 2.425\text{AU}]$. Moreover, it turns out that the semi-major axis - diameter plot excludes some asteroids which do not have an estimate of their size known (in particular cannot be extracted from the Small Body Database). In particular, in the realm of small a, Figure 12 does not show 38070 Redwine at $a = 2.143\text{AU}$, and on the other extremity it omits a nameless asteroid 97276 at $a = 2.482$. A possibility that if the sizes of these objects were known, or if more objects were included in the database, the missing corners of the plot on Figure 12 would also be filled with scatter points, disproving a hypothesis of the linear trend observed here, cannot be excluded. Therefore, the line shown on Figure 12 should be considered an observation and conjecture only, and could only be moved to the realm of factual correlations if in the future a similar behaviour were observed on a plot comprising all or almost all the asteroids, with their taxonomies and diameters being well documented.

- **Outliers.** While almost all asteroids in the database belong in the area of $\sim a > 1.6\text{AU} \wedge d < 300\text{km}$, several objects transcend those boundaries:
  - V class representative 4 Vesta, with a diameter of 525.4km (second largest asteroid in the Solar System), is the only V class representative outside of the cluster described above, and contrary to all othe V class objects in this dataset, it is huge
  - B class object 2 Pallas has a diameter of 513km, only slightly smaller than Vesta
  - Xk class Near Earth Asteroid 3103 Eger has a diameter of 1.5km and semi-major axis of 1.404AU
  - Sa class NEA 25143 Itokawa, with a diameter of 330m, can be found at $a = 1.324\text{AU}$. The interest in that asteroid is justifiably high, since the Hayabusa mission (Yano *et al.* 2006) has provided us with spectra registered from proximity and a small material sample brought back to Earth
  - B-class NEA 2100 Ra-Shalom is the only inner object in the database, with $a = 0.832\text{AU}$
- **Disjoint ranges of a for classes V and D.** Since the only class with a high number o representatives other than V available in this dataset is D (see Figure 10), it is the only other class for which more defined trends or clusters could be expected to be found. However, the points seem to be scattered quite uniformly in the region of $a \in\sim [2.5\text{AU}, 3.25\text{AU}]$, $d \in\sim [10\text{km}, 200\text{km}]$, $e \in\sim [0, 0.3]$ (with the smallest one being 2606 Odessa of diameter 15.91km at $a = 2.759\text{AU}$ and the largest being 375 Ursula with $d = 216\text{km}$ and $a = 3.127\text{AU}$). It seems doubtful that any linear trend among this data could possibly be apparent and statistically significant. The only firm

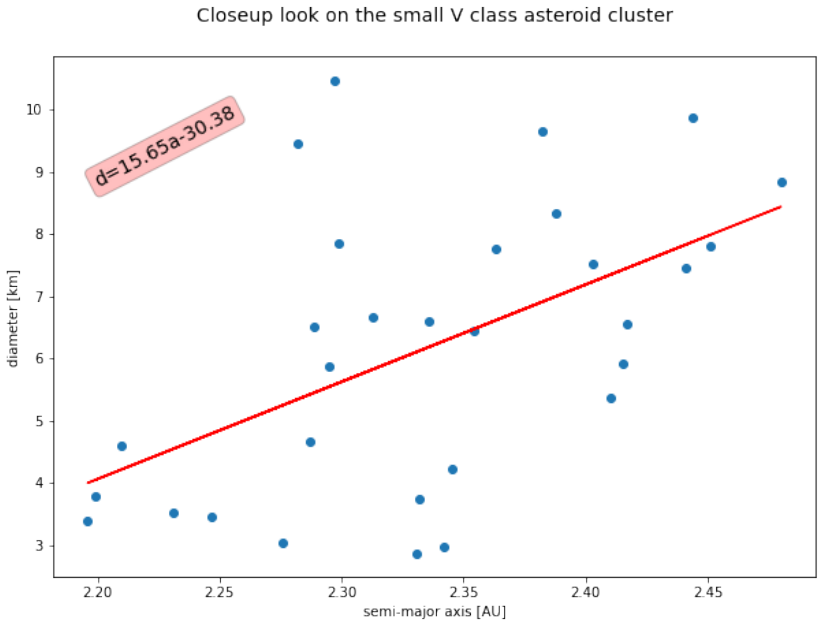

Figure 12: A closeup look at the cluster of small V class objects in the IRTF dataset. A linear trend, which seems to be hinted at on Figure 11 can be found and fitted for, but does not look extremely strong and may turn out to be coincidental if a much broader dataset were considered.

conclusion is that the V class and D class objects occupy almost entirely disjoint ranges in terms of the the semi-major axis parameter. Indeed, as mentioned above the maximum a among the V class in this dataset falls at $a = 2.482$, whereas in the D class the shortest semi-major axis representative is 284 Amalia at $a = 2.359$AU, but the second shortest already exceeds 2.5AU (785 Zwetana, with $a = 2.572$).

- **Big sizes of X-spectrum objects** One may notice that all asteroids from the X-spectrum (aside of the outlier 3103 Eger), seem to avoid the area of small sizes. Indeed, in the regime of $a > 2$AU, the smallest object, 678 Fredegundis, has a diameter of almost 40km, while the lower bounds on C-complex and S-complex objects at $a > 2$AU in the same database fall near 5AU and 10AU, respectively (see Table 3). However, there are generally only 15 X-complex objects in this database, so the observation may again be misleading. Indeed, much smaller X-complex objects can be found in both Primass and SMASS databases (2.3,2.4), so the tendency seems to be only a dataset - specific feature.

	End members	C-complex	S-complex	X-complex
smallest asteroid	27343 Deannashea	26760 (nameless)	2504 Gaviola	678 Fredegundis
diameter	2.864 km	5.4 km	10.579 km	**39.585** km

Table 3: Lower bounds on sizes of objects from different taxonomic categories in the IRTF database. The relatively large value for X-complex is shown in bold.

As mentioned in the beginning, similar classifications for Primass and SMASS are less reliable and did not exhibit unique unexpected clusters. Only the $a - d$ parameter plots will be shown here in Figure 13, to show the general range of semi-major axes and sizes present in those surveys, refraining from further discussion.

6 Detecting and measuring absorption bands

One of the big objectives of this project was to automatize the process of detecting and measuring absorption bands in reflectance spectra of asteroids. Extensive reliance on methods of the CANA library (De Pra *et al.* 2018), as in the case o taxonomic classification (section 5), was not possible in this case, as the tools implemented for measuring absorption bands in the CANA library are oriented towards low dataset size manual analysis approach, where the

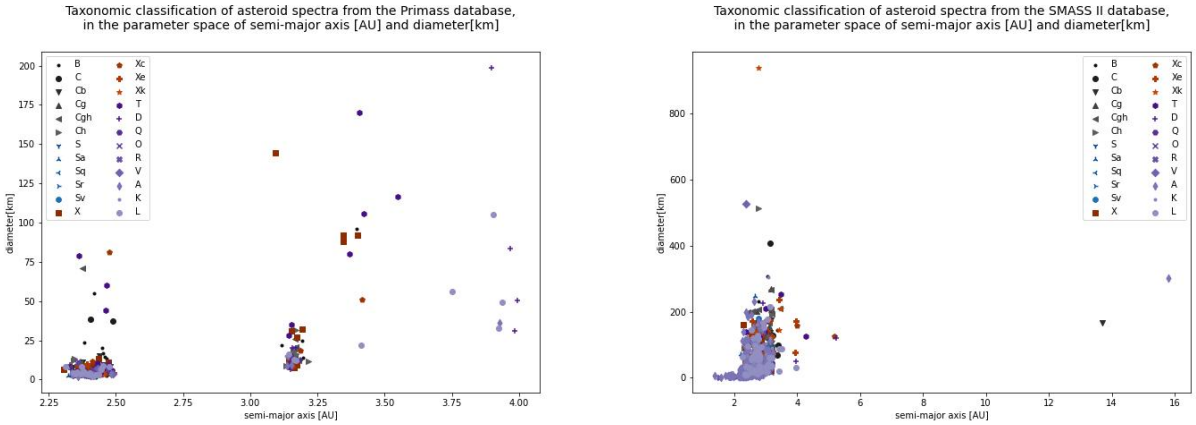

Figure 13: $a - d$ parameter space classification plots for Primass (**left**) and SMASS II (**right**). The plot for the first phase of SMASS is not included as it looks very similar to the one from the second phase. We can see that the Primass survey included bodies under 200km only, with a focus on very small bodies $\sim<$ 25km, clustered in two groups ($a \in\sim [2.3\text{AU}, 2.5\text{AU}]$ and $a \in\sim [3.1\text{AU}, 3.2\text{AU}]$). No additional clusterings can be found in the regime of eccentricities or inclinations. The aggregation of similar semi-major axis is due to the fact that the Primass survey targets specific asteroid families. Distribution of semi-major axes is closer to uniform in the regime of $a \in\sim [2\text{AU}, 4\text{AU}]$ or SMASS II, with two outliers from the centaur family: Cb class object 2060 Chiron at $a = 13.71\text{AU}$ and A class (assuming the classification with CANA was not misleading, as discussed in 5.3 A seems to be a common false match and indeed in the Small Body Database (N. JetPropulsionLaboratory n.d.(a)) the same object is classified as spectral type D) 10199 Chariklo

range of wavelengths corresponding to a candidate absorption band needs to be defined by the user. Therefore, a novel approach needed to be invented, and numerous challenges were faced in the process. Ultimately, although the highest degree of automation desired proved to be beyond reach, a satisfactory algorithm was developed allowing to quickly analyze an arbitrary number of spectra provided that they belong to a class with known characteristics. This section outlines the methodology developed and exemplifies it's application to the data described in section 2.

6.1 Preprocessing

The general idea for automatic detection of an absorption band is to look for such a range o wavelengths $[w_{\min}, w_{\max}]$ in a spectrum, that

$$\forall_{w \in (w_{\min}, w_{\max})} r(w) < cont(w)$$

where $r(w)$ is the reflectance value for a given wavelength, and $cont(w)$ is the *continuum* line, which is meant to approximate the spectrum without the presence of the absorption bands and is a linear function between the spectrum at the ends of the interval:

$$\forall_{w \in [w_{\min}, w_{\max}]} cont(w) = r(w_{\min}) + \frac{w - w_{\min}}{w_{\max} - w_{\min}} \cdot (r(w_{\max}) - r(w_{\min}))$$

In their default form, spectra downloaded from databases described in section 2 were not suitable for automatic detection, due to high level of noise in the data. Indeed, in order to search for a suitable interval $[w_{\min}, w_{\max}]$ in an automated way, we would need consistency in the moment when modifying this range in a semi-continuous manner would result in a change of a criterion:

$$V(w_{\min}, w_{\max}) = \begin{cases} 1 \text{ if } \forall_{w \in (w_{\min}, w_{\max})} r(w) < cont(w) \\ 0 \text{ otherwise} \end{cases}$$

i. e. for two similar spectra we would want that if $w_{\min}, w_{\max}$ are the boundaries of the absorption band range present in both, then they would both satisfy $V(w'_{\min}, w'_{\max}) = 1$ for all, or at least most pairs of $w'_{\min}, w'_{\max}$ such that $[w'_{\min}, w'_{\max}] \subseteq [w_{\min}, w_{\max}]$, and $V(w'_{\min}, w'_{\max}) = 0$ for all $w'_{\min}, w'_{\max}$ such that $[w'_{\min}, w'_{\max}] \nsubseteq [w_{\min}, w_{\max}]$. Spectra with a high, or even moderate level of noise, do not satisfy such a requirement, as local noise-induced variations would result in many subranges $[w'_{\min}, w'_{\max}]$ in which the value of $V(w'_{\min}, w'_{\max})$ would flip between 0 and 1 reacting to

minute changes of range boundaries (due to the fact that noise-induced small peaks in reflectance would rise above continuum lines drawn between nearby points in the reflectance spectrum plot).

One could consider different ways of alleviating this problem by preprocessing a spectrum so as to remove the noise by generating a smoother proxy of the original spectrum. Examples of ideas would be rebinning the spectrum to decrease spectral resolution or fitting a polynomial function or a B-spline type of function. Both of these are readily implemented in the CANA library (De Pra *et al.* 2018). However, the method chosen and implemented in this project is different, and it relies on average smoothing of the reflectance spectrum, which has an advantage of being more faithful to the original data and being able to preserve some of the low-scale features which may have physical meaning and could be lost in a polynomial fit. It has been implemented using a *Pandas* ***rolling*** function. In it's traditional form, average smoothing would take a window of size n, then convolve the spectrum with an array o length n and all values equal to $\frac{1}{n}$, so as to arrive at:

$$\tilde{r}(w_i) = \frac{\sum_{j=-\lfloor \frac{n}{2} \rfloor}^{-\lfloor \frac{n}{2} \rfloor + n - 1} r\left(w_{i+j}\right)}{n}$$

where $w_1, w_2, ...$ are consecutive wavelengths for which reflectances are provided in the spectrum, so that fro w_i we are averaging over values $w_{i-\lfloor \frac{n}{2} \rfloor}, ..., w_{i+\lfloor \frac{n}{2} \rfloor}$ (or one of the end values excluded if $2|n$). This approach, however, would require that a wavelength for which a new proxy spectrum $\tilde{r}$ were computed, would have $\lfloor \frac{n}{2} \rfloor$ other data points between itself and the beginning or end of the wavelength range covered in the dataset, so the "tails" of the spectrum would not be preserved and the wavelength range covered would shrink. It is obviously not desired and could be detrimental for the objective of this work, especially if an absorption band we are trying to detect were located close to the beginning or end of the wavelength range, as is the case for example for the first pyroxene/olivine absorption band[1] for some of the spectra in the IRTF database (see section 9). Therefore, parameters of the function were altered so as to compute:

$$\tilde{r}(w_i) = \frac{\sum_{j=\max(-\lfloor \frac{n}{2} \rfloor, 1-i)}^{\min(-\lfloor \frac{n}{2} \rfloor + n - 1, N-i)} r\left(w_{i+j}\right)}{n}$$

where N is the length of the array of wavelengths at which the spectrum is sampled, so that a formula is valid and can be evaluated for all $i \in \{1, 2, ..., n\}$.

This procedure is applied in the code for all spectra before subjecting them to the absorption band analysis algorithm (6.2). Window size can be defined by the user. In this study, two values of n were experimented with: $n = 11$ and $n = 29$ corresponding to two extremities of the precision/smoothness trade-off (originally $n = 10$ and $n = 30$ were worked with, however, an odd-sized window seems favorable as it is perfectly symmetrical around a given wavelength, whereas an even-sized window could have subtle biases related to the extra point on one side of the window only). An example of how the averaging procedure affects the original spectrum, and what kind of subtle, yet significant differences arise from a different choice of the averaging window, is shown on Figure 14.

6.2 The algorithm

This and next section will describe the algorithm developed to detect absorption bands in a spectrum. As explained in 6.1, the first essential step was to smoothen the spectrum by applying average smoothing of a chosen window size. The algorithm deals with such a smoothened spectrum, therefore wherever application of the algorithm to a reflectance spectrum will be referred to, the application of average smoothing before running the algorithm should be assumed, even if not stated explicitly.

The first approach towards automatic absorption band detection attempted in this project was to look for peaks in the signal as candidates for $w_{\min}$ and $w_{\max}$ for absorption bands, and for peaks in the negative of the signal ($-r(w)$) as candidates for band centers, then to refine the interval boundaries corresponding to each band until the condition

[1] Pyroxene/olivine absorption band, in this case, refers actually to an absorption feature which is composed out of three olivine absorption bands and one pyroxene absorption band. Even though the term absorption band should most formally be reserved for one band which could be directly linked with a physical process which is responsible for that band by absorbing light of specific energy for a specific purpose, whereas a visual dip in the spectrum which may need to be decomposed into several bands to assign it a physical explanation, should rather be called a feature, the two terms tend to be used interchangably and the meaning of word *band* in a given sentence can always be inferred from the concept. Therefore, the section 6 will repeatedly use the term *band*, even though in most cases only a visually interpreted absorption *feature* is being implied.

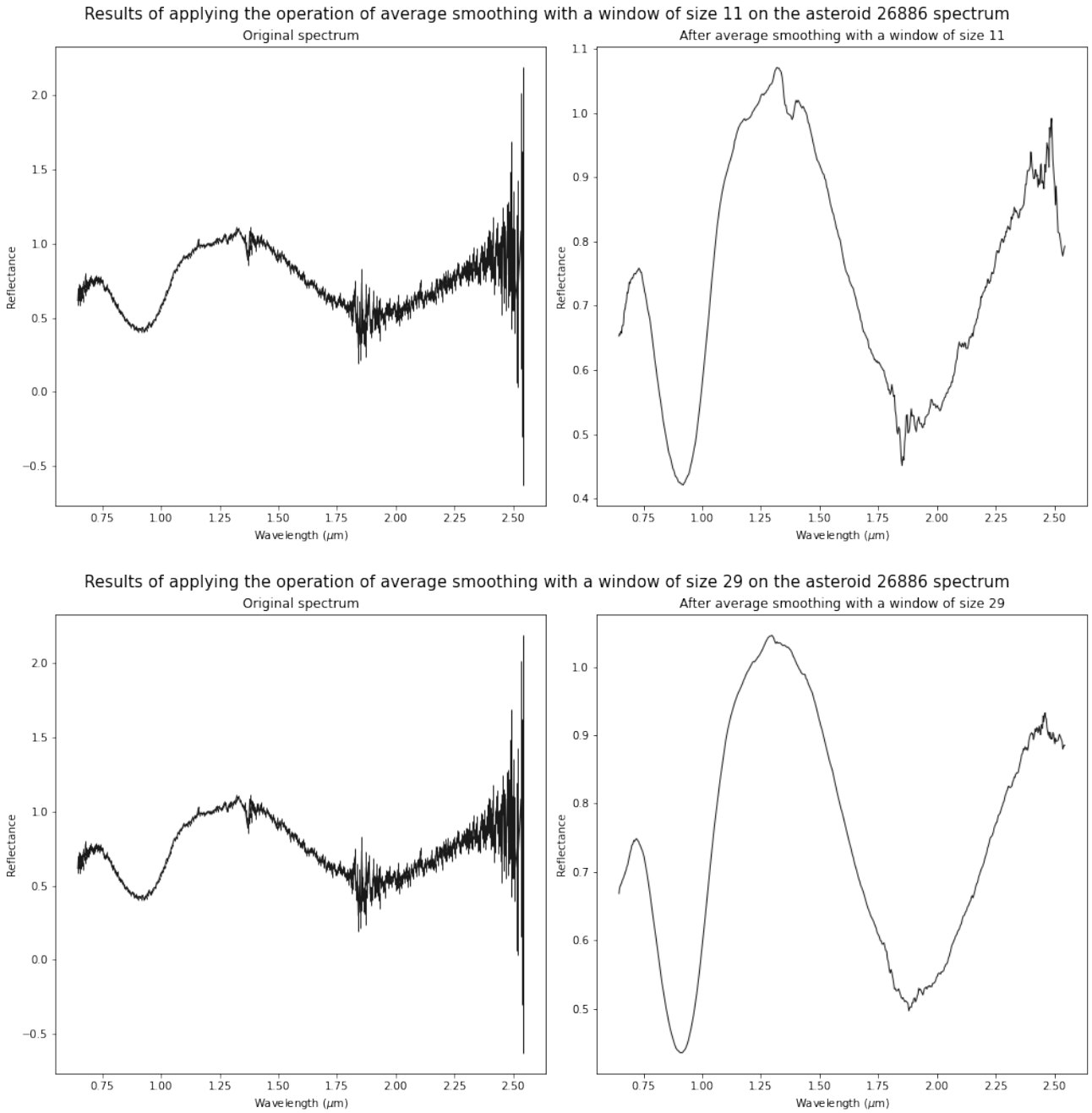

Figure 14: Example of the effect of average smoothing with two different window sizes, performed on a reflectance spectrum of a V class asteroid 26886 from the IRTF database (2.5). **Upper:** window size of 11 preserves more detail of the original spectrum, such as the small feature around 1.4μm which may be physical. On the other hand, the lack of smoothness due to effects of the noise persists in some areas, especially for longer wavelengths where measurement errors had been huge. **Lower:** using a window of size 29 results in a plot which is smooth at a sufficient level for virtually any application. However, some of the finer detail is irrevocably lost. Notice in particular, that the leftmost peak in the spectrum ($\sim 0.75\mu$m) is slightly lower on the lower graph, due to the fact that it has been averaged together with more of the lower values after the peak. In fact, the same holds true for the other two major peaks ($\sim 1.3\mu$m and $\sim 2.45\mu$m), but is less striking because the areas around these peaks have less steep slopes. Huge uncertainties on the very end of the original spectrum lead to the fact that some lack of smoothness persists near 2.5μm even in the lower graph corresponding to the window 29, this is rather anomalous, but as it is the very end of the spectrum, will not have significant impact on the evaluation inferred from absorption band analysis. Both figures were generated with a function utilizing a CANA library (De Pra *et al.* 2018) **plot** function. The plots are colorless and not particularly elegant, which is why the function was not used for most data presentation purposes in this project. However, it was a convenient tool to quickly evaluate the result of applying different operations (in this case, average smoothing with two different window sizes) to an original spectrum. The original spectrum shown on the left of the upper and lower figure is exactly the same.

$V(w_{\min}, w_{\max}) = 1$ would be satisfied. This seemed to be the only viable approach with which we could identify all absorption bands in all kinds of spectra, without any prior knowledge as of what to expect. Unfortunately, this ambitious goal could not be met - no kind of tinkering with the code, nor adding additional parameters, could prevent the artifacts such as seen on the left o Figure 15 from arising frequently.

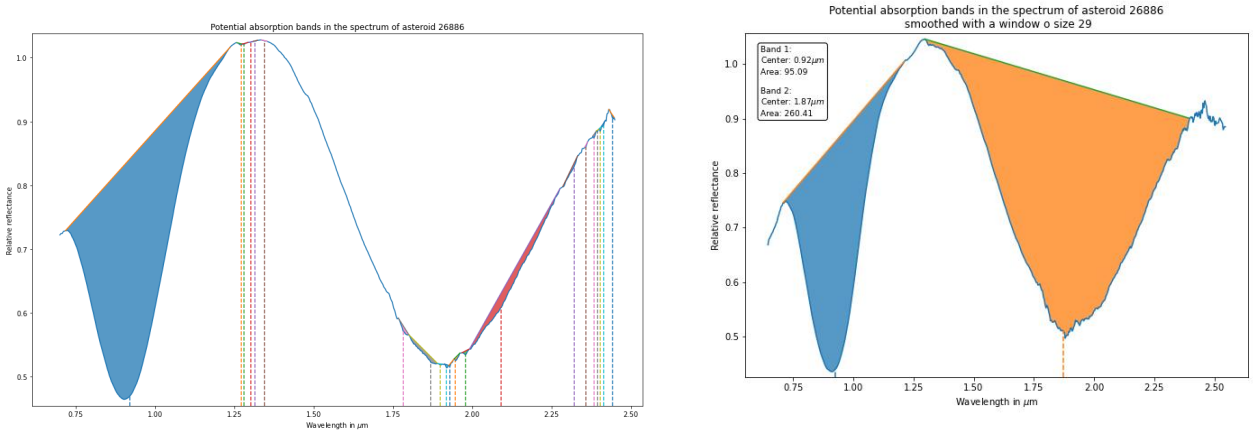

Figure 15: **Left:** screen capture from the old version of a project, with a result of the malfunctioning automatic absorption band detection algorithm for an example asteroid 26886 (V class asteroid from the IRTF database, same as on Figure 14). **Right:** two absorption bands found by the new algorithm (6.2), which yielded consistent and intuitive results for all V class asteroids in the IRTF database. The old algorithm had been able to accurately identify the first band, but then it identified a myriad of tiny bands of no physical meaning where a large single band should be identified instead.

This lead to a conclusion that even smoothened spectra cannot be analyzed to detect absorption bands without providing any *a priori* information - even a relatively featureless smooth curve resultant from an average smoothing procedure has minute details which can become false positives for absorption bands. Presumably, if far more attention were devoted and a variety of special cases were included in the algorithm, or if a machine learning oriented approach were used to differentiate between genuine and fake absorption band candidates based on a comparison with a human-labeled training set, this approach could be refined until it would be able to accurately detect absorption features in a majority of cases.

However, in this project, the entire approach has been abandoned in favor of a somewhat different objective which has been deemed superior due to it's potential of almost 100% accuracy. Specifically, we should realise that positions and widths of absorption bands are not random, but can usually be well predicted (on the first approximation level) based on existent knowledge about reflectance spectra of asteroids. In particular, majority of reflectance spectra which do have noticeable absorption features, have the two features which correspond to the two most commonly encountered classes of minerals: $\sim 1\mu$m feature due to both olivines and pyroxenes and a $\sim 2\mu$m feature due to pyroxenes only. Even if it is not the case, when studying a certain dataset the same kind of absorption bands are expected to appear in numerous cases, and a comparison between features in the same region of a spectrum is an interest of comparison (see for example the case of M type asteroid features sought for in the Fornasier survey described in 2.1 and evaluated in 9). Therefore, it is reasonable to first delineate a subset in the bulk of data for which a consistent type of absorption feature should be identifiable, then define proxies for band centers and widths which we could expect, and only after run an algorithm which would find each feature in every spectrum of that subset with high precision. This is the core assumption behind the new algorithm.

The design of the algorithm is mostly inspired with an idea to mimic the way a human would try to find the most accurate range of wavelengths to represent an absorption band, but taking into account that the computer is "blind" and cannot immediately guess the right range, but can rather start out by finding one range that satisfies the desired condition and then keep modifying it in until the biggest satisfying interval is found. The other essential inspiration was to imitate the clockwork of a binary search algorithm, so that instead of modifying the interval step by step in a linear fashion, repeated multiplication of the modification step by a constant would grant the process more of a logarithmic behavior. Except in this case the incentive to use binary-search-like methodology was not solely driven by

a goal of reducing computational complexity, but most importantly by the goal of reducing a likelihood that a rare case of $V(w'_{\min}, w'_{\max}) = 0$ for some $[w'_{\min}, w'_{\max}] \subset [w_{\min}, w_{\max}]$ would stop the execution in a wrong place and lead to identification of a smaller absorption band than desired. Indeed, such a situation cannot be fully prevented, even taking the average smoothing process into account. On one hand, it might arise due to some subtle fluctuation induced by residual impact of noise. Remember that some trace of the noise may be retained even after average smoothing and that the choice of the smoothing window is constrained by a trade-off between smoothness and faithfulness to the original data. Hence, a subtle bump of no physical meaning appearing here and there will inescapably remain plausible. On the other hand, such discontinuity of the $V(w'_{\min}, w'_{\max})$ condition may have physical origins coming from the complexity of the feature. Suppose that in terms of visual properties, a band can be described as a big dent in the curve, but there is a smaller dent inside of it. This could well be due to a blend of two absorption bands of different widths. However, the algorithm as it is designed will not be able to detect them simultaneously, it would rather have to be run twice with different parameters, to detect the bigger and the smaller one, separately. While trying to detect the bigger one, it will be prone to failure as it would get stuck exploring the wavelength range on the border of the small band inside the band, even though it should actually explore much further. Measures taken in the algorithm, which are inspired by the binary search methodology, allow to suppress such misfits of the wavelength range in practically every case.

The initial idea needed to be refined in numerous ways before satisfactory performance was achieved. A detailed outline of the final version of the algorithm is presented as a flowchart on Figure 16. Bear in mind that the flowchart illustrates a run for one absorption band of interest defined by provided initial proxies of w_{cen} - center of the band and Δw - width o the band. The full absorption band search function would take a list of those values for each band we are trying to find. Default parameters are $\mathbf{w}_{\text{cen,list}} = [w_{\text{cen1}}, w_{\text{cen,2}}] = [1\mu\text{m}, 2\mu\text{m}]$ and $\mathbf{\Delta w}_{\text{list}} = [(\Delta w)_1, (\Delta w)_2] = [0.2\mu\text{m}, 0.2\mu\text{m}]$, corresponding to the olivine/pyroxene 1μm and pyroxene 2μm absorption bands. As long as such bands really are common in the subset of spectra taken under study with the function, initial estimates do not need to be very precise, especially for the widths Δw, and the rate of success should still be very high.

To facilitate the understanding of the algorithm, let's also outline the major components of an iteration over one candidate here:

1. Initial proxies w_{cen}, Δw, as defined by the user, are loaded and converted into $w_{\min}$,$w_{\max}$.
2. A check is made if such interval inferred from initial parameters satisfies the condition $V(w_{\min}, w_{\max}) = 1$:
   (a) If yes, the interval is progressively expanded in a symmetric manner by moving the left end $\frac{\Delta w}{2}$ to the left and the right end $\frac{\Delta w}{2}$ to the right, until the condition of $V = 1$ no longer holds, then a step back is taken to the last case where the condition did hold.
   (b) If not, the interval is progressively shrunk by multiplying it's length by 0.75 (and keeping it symmetric around w_{cen}), until the condition is met. It has to be satisfied at some point, because in the program the wavelength range is discretized so if we keep reducing the interval size in an exponential manner, eventually both ends will correspond to the same wavelength in the wavelength array, and $\forall_w V(w, w) = 1$ because an open interval of the same two ends is an empty set: $(v, v) = \emptyset$ (so the condition for $V = 1$ is met automatically). When the condition is met, a check is made to verify if the width of the satisfying interval $w_{\max} - w_{\min}$, is above a predifined threshold:
      i. If yes, the algorithm will carry on with the interval found.
      ii. If not, it means that the shrinking has converged to a tiny interval which cannot be interpreted as an absorption band (or perhaps it has even converged to a single point, as just explained). In that case the conclusion is made that the band of desired proxy center and width is not present in the spectrum. The program would then move on to study another candidate w_{cen}, Δw, if there is any pair provided by the user (or inferred from default parameters corresponding to olivine and pyroxene features) left.
3. The algorithm moves on to an attempt of broadening the interval found in a "little by little" manner, checking first how much it can move to the left without violating the $V = 1$ condition, then to the right, always jumping in steps of $\delta w = \frac{w_{\text{max,p}} - w_{\text{min,p}}}{4}$, where $w_{\text{max,p}}$ and $w_{\text{min,p}}$ are ends of the interval found in the previous step. It then moves on to halve the step δw and do the same search again, until the step becomes comparable to the spectral resolution of the spectrum under study ((average) wavelength separation between two consecutive entries in the wavelength array). Moving at steps lower than that resolution would obviously make no sense, since the wavelengths drawn from the wavelength array would no longer change. The new largest interval satisfying the condition $V = 1$ is saved.

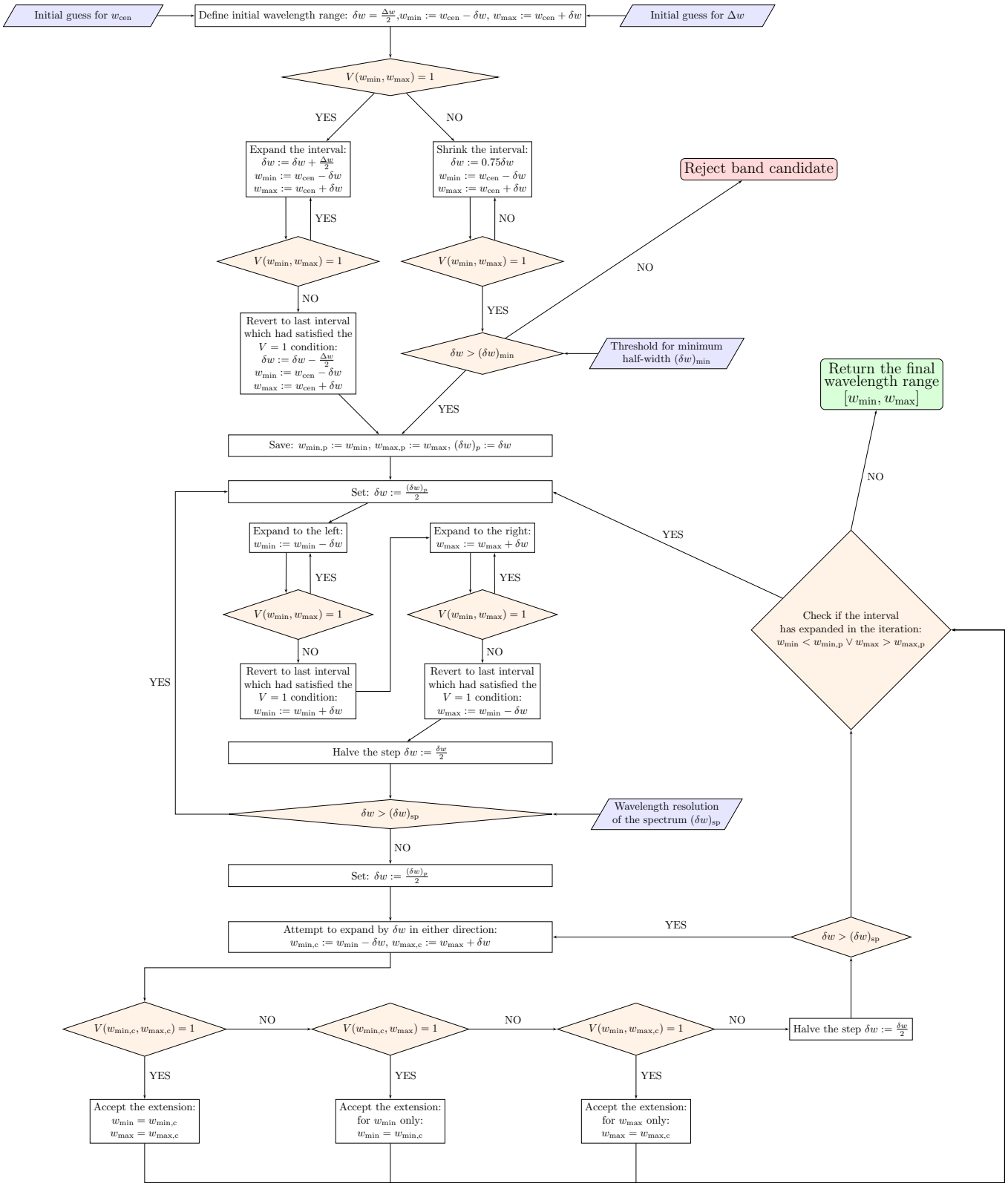

Figure 16: Flowchart presenting the absorption band identification algorithm as implemented in the final version of the code (Morawski 2023-2024). Actual implementation in the code is slightly more complicated as we cannot work with wavelengths directly but with indices of the wavelength array. As explained in the text, it essentially consists of fitting the first proxy of an interval which satisfies the condition $V(w_{\mathrm{min}}, w_{\mathrm{max}}) = 1$, then looping over trying to add corrections of varying size (from a distribution of exponentially decreasing sizes) on the left, right, or both sides of the interval, until no change is made in a given loop. If the very first attempt at finding an interval which satisfies the $V = 1$ condition fails, the search is inconclusive and no band candidate is returned for given w_{cen} and Δw. Otherwise, once the first valid interval is returned, the algorithm ensures to keep a valid candidate so that the output wavelength range would satisfy the condition also.

4. δw is set back to $\frac{w_{\mathrm{max,p}}-w_{\mathrm{min,p}}}{4}$, and an attempt is made to expand the interval in at least one direction, or possibly both, by consecutively checking smaller corrections of δw in the same manner as in step 3 (that is halving at each iteration, which in both cases is a tribute to the binary search algorithm inspiration mentioned above).

5. Once an extension is made successfully, or δw drops below the spectral resolution again, a check is made to see whether the interval has expanded in any way compared to the $[w_{\mathrm{min,p}}, w_{\mathrm{max,p}}]$ interval with which we were starting out with in step 3:

   (a) If yes, the algorithm moves back to step 3, updating values of $w_{\mathrm{min,p}}, w_{\mathrm{max,p}}$ to the new candidate interval.

   (b) If not, a conclusion is made that the final range has been found, and the algorithm concludes. Strictly speaking, this concludes the part of the algorithm which is shown on Figure 16, but the function as written in the code (Morawski 2023-2024) moves on to carry out additional steps, which will be elaborated on in the following paragraphs.

Before moving on to discussion of how a successfully found absorption band is then treated by the program, let's conclude this algorithm description with one more remark. One may wonder, why are the steps 3 and 4 mentioned above both necessary, and why is it necessary to loop over them multiple times. As it turned out during the development of the algorithm, if only one way of expanding the interval is introduced, or if the correction δw were to be only reduced in size and the algorithm were to conclude on the first instance of it reaching the lower bound on it's size (the spectral resolution), the algorithm is prone to get stuck on some local minima, and fail to take into account the fact that as one side of the band has successfully been pushed to the left, the same modification to the right end which had previously lead to $V = 0$ can now be done while retaining $V = 1$. Only a fusion of both expansion techniques, and repeated execution until no further expansion has been achieved in either, lead to consistent detection of absorption feature in the same way as a human would draw them based on intuition, whereas any other architecture of this part of the algorithm lead to suboptimal results for some of the spectra on which it was tested.

Having identified an absorption band as a feature in an interval $[w_{\mathrm{min}}, w_{\mathrm{max}}]$, the code moves on to do additional analysis steps for the band:

1. Identify a first candidate on the band center $w_{\mathrm{cen},0}$ as such wavelength that $\forall_{w\in[w_{\mathrm{min}},w_{\mathrm{max}}]} cont(w) - r(w) \leqslant cont(w_{\mathrm{cen},0}) - r(w_{\mathrm{cen},0})$, i. e. the wavelength where maximum absorption with respect to the continuum line spanned between w_{min} and w_{max} can be measured

2. If the maximum absorption with respect to the continuum, $cont(w_{\mathrm{cen},0}) - r(w_{\mathrm{cen},0})$, turns out to be lower than a threshold defined by the user (but set to a very low value of 0.01 by default), then a candidate absorption band is disregarded. This "post-production" removal step is aimed at sieving out cases where a very shallow dent in the spectrum which is likely to have no physical meaning, were found and identified as an absorption band by the algorithm. It is only needed in rare cases.

3. Find a set $W_c \subset [w_{\mathrm{min}}, w_{\mathrm{max}}]$ of all wavelengths w_c such that $1 - \frac{cont(w_c)-r(w_c)}{cont(w_{\mathrm{cen},0})-r(w_{\mathrm{cen},0})} < \xi$, where the threshold ξ on similarity to the absorption maximizing wavelength can be defined by the user and is by default equal to 0.05.

4. Define the final fit for the center of the line as $w_{\mathrm{cen,final}} = \langle W_c \rangle$, i. e. as the arithmetic mean of all wavelengths in the set W_c. Steps 3 and 4 are aimed at altering the choice of w_{cen} from the first estimate of a sole wavelength which minimizes absorption from the continuum, so as to mitigate possible distortions due to traces of noise in the smoothened reflectance curve.

5. Calculate the band area as $A = \int_{w_{\mathrm{min}}}^{w_{\mathrm{max}}} (cont(w) - r(w))dw$.

6. Save the band and information about w_{cen} and A to an object.

Based on the information gathered during such analysis, plots such as the right side of Figure 15 can then be generated. Unlike the outcome of the old algorithm, the result of the new one matches extremely well a delineation of absorption bands as it would be done by a human data analyst. In section 9, more examples will be presented, as we move on from the description of the algorithm itself to it's application to the data from databases introduced in section 2.

7 Corrections for temperature and phase angle

7.1 Motivation

An important notion in asteroid reflectance spectra analysis, especially in the context of inferring mineralogical compositions based on empirical formulas (which are based on laboratory analysis of meteorite spectra), is that a number of variables affect the spectra of remotely sensed asteroids and need to be accounted for in order to be able to compare the results. As summarized in V. Reddy *et al.* 2015, these variables are:

- **Grain size.** A spectrum of an imaginary asteroid with the same mineralogical composition would look differently depending on the typical scattering particle (grain) size on it's surface. Larger particle sizes typically lead to deeper bands, bluer (more negative) spectral slope and lower reflectances in general. However, no statements have been made regarding potential influence of this factor on band center positions nor band area ratio, and no empirical formulas have been proposed to correct for this effect. We shall therefore assume that this effect is negligible for the purposes of the analysis presented here. Indeed, it would only be relevant for precise calculation of spectral slopes, which is not a mattrer of focus here. Moreover, a righteous way of correcting for this effect is not even attainable - it would require information about average particle sizes on asteroids, which is not available except for a few cases where space mission sample collection or fly-by observation allows for making an estimate.

- **Temeprature.** Mineralogical formulas are derived based on laboratory analysis of meteorite spectra, which are measured at room temperature ($\sim$ 300K). Temperature affects wavelength ranges of Fe^{2+} crystal field absorptions in olivines and pyroxenes (Singer and Roush 1985), which in turn influences characteristics of the spectra such as band areas and band II center position. Therefore, a correction must be introduced to apply lab-derived formulas to spectra of asteroids, which obviously have different (in most cases much lower) surface temperatures then 300K. We should remember, however, that it is not the corrected, but the original band position and area, what should be considered a "true" position or area. Instead, a correction should be thought of as an artificial way of bringing asteroids to the lab and asking how the measured properties of the spectrum would have changed if they had room temperature.

- **Phase angle.** It is well known that any observation in the solar system needs to take into account a phase factor $\Phi(\alpha) \in [0, 1]$ multiplying the light intensity calculated based on object's distances form the Sun and Earth only, where α is the phase angle, i.e. the Sun-target-Earth angle (e. g. Lumme, Bowell, and Harris 1984). However, it is usually spoken of in terms of bolometric magnitude, so the dependence on the wavelength is not being discussed. As research of asteroid spectra has shown, when considering brightness change due to phase angle as a function of wavelength $\Phi_\lambda(\alpha)$, it turns out that the effect is stronger for shorter wavelengths (higher $\left|\frac{d\Phi_\lambda(\alpha)}{d\alpha}\right|$ when λ is smaller), and conversely, spectra of the same asteroid observed at a higher phase angle would experience a stronger measured reflectance drop in the blue part of the spectrum then in the red part, an effect referred to as *phase reddening*. It should not be surprising that such uneven treatment of different parts of the spectrum by the phase angle influence would affect the band area ratio and should be corrected for. Unlike the temperature correction, in this context it is rather the values after the correction which should be thought of as "real", since the correction removes an effect induced by the geometry of an observation and brings us to the values which would have been obtained from evaluating the same asteroid observed without the phase effect (at opposition).

Means to include temperature and phase angle corrections in this project needed to be considered for the sake of performing a reliable analysis of the IRTF spectra absorption features in the mineralogical context (9.2).

7.2 Empirical formulas for the corrections

Regarding the temperature, Sanchez *et al.* 2012 studied meteorites at different lab temperatures (between 80 and 400K) and derived linear fits for corrections which should be made to band area ratios and band II center positions for asteroids of given temperature:

$$\Delta\text{BAR} = 0.00075T - 0.23$$

$$\Delta\text{BII} = 0.06 - 0.0002T$$

where BAR is the band area ratio (band II area divided by band I area), BII is the band II center position expressed in μm, T is the temperature of an asteroid in Kelvins and Δ symbolizes a change of a given spectral properties, so the corrections above need to be added to BAR and BII values inferred from the spectra. It is important to point out that the study used ordinary chondrite meteorites only, so the formulas may not be optimally suited for all asteroids.

However, no other formulas for the band area ratio have been proposed, so the Sanchez *et al.* 2012 formula for ΔBAR will be used in the entirety of the analysis in this section. Regarding band positions, alternative formulas fitted for howardites and eucrites (which are structurally closer to V class asteroids) are proposed in Vishnu Reddy *et al.* 2012:

$$\Delta\text{BI} = 0.01656 - 0.0000552T$$

$$\Delta\text{BII} = 0.05067 - 0.00017T$$

Notice that the two formulas for BII correction are very similar to each other, and the formula for BI correction yields results of the order of 0.01μm or less, whereas the Sanchez *et al.* 2012 approach assumes no correction applied to BI. Hence, a choice of one or the other set of formulas should not affect the result drastically. In this project, Sanchez *et al.* 2012 formulas for band center corrections are used when dealing with S-complex asteroids, as they are often parent bodies of ordinary chondrites, and Vishnu Reddy *et al.* 2012 formulas otherwise (R and V classes), with the awareness that it may not always be a fully accurate choice considering each asteroids' individual composition, but with a reasoning that due to the subtelty of the difference between both approaches pointed out here the effect of that choice on overall results should not be drastic (in particular, the analysis of V class object in 9.2.4 is of rather statistical nature, whereas the analysis of Sa class objects in 9.2.3 focuses on two objects the knowledge on the characteristic of which is well established).

As for the phase angle, the only spectral parameter with an established formula for the correction is band area ratio, for which Vishnu Reddy *et al.* 2012 fit:

$$\Delta\text{BAR} = -0.0292 \cdot \alpha$$

where α is expressed in degrees. Although it will be used in this study, there are reasons to believe that it may not always be accurate and further scientific investigations need to be conducted on the topic of phase angle corrections before corrections for the phase reddening can be applied reliably to every case:

- The formula has been derived based on observations of one asteroid only (4 Vesta)
- It is considered to apply in a range of $\alpha \in [0°, 25°]$ only.
- The correction is quite large and it is clear that the linear relationship is not likely to extrapolate for bigger angles - already at an angle of 35° it would lead to $|\Delta\text{BAR}| > 1$, which could push BAR to negative numbers in many cases
- Sanchez *et al.* 2012 did not find evidence that such a strong correction would be needed for ordinary chondrites

Taking all of these into account, in this project the formula will be used for phase angles under 25°, an approximation with the edge case value of $-0.0292 \cdot 25$ will be used for bigger phase angles, and mineralogical composition results derived with formulas introduced in section 8 will be given less credibility if corrected band area ratios will be negative.

7.3 Approximating asteroid temperature

Having discussed how these two corrections work, let's move on to evaluation of possibilities to apply them to the IRTF data within the framework of the project. In order to estimate surface temperature, an assumption of gray body equilibrium temperature is made. It states that the solar radation energy received by the asteroid should be equal to the energy it emits (mostly in the infrared). The flux of energy received is equal to:

$$(1 - A)\pi r^2 \frac{L_\odot}{4\pi d^2}$$

where A is the albedo of the asteroid (so the A fraction of the solar radiation energy is reflected instead of being absorbed), r is the radius of the asteroid (approximated as a sphere), $L_\odot = 3.827 \cdot 10^{26}$W is the solar luminosity and d is the distance from the asteroid to the Sun. The power at which the body is emiting is given by:

$$\epsilon\sigma T^4 4\pi r^2$$

where ϵ is the emmisivity (correction factor for the asteroid not emiting like a perfect black body) and $\sigma = 5.67 \cdot 10^{\breve{}8}\text{Js}^{\breve{}1}\text{m}^{\breve{}2}K^{\breve{}4}$ is the Stephan-Boltzman constant. Matching the two expressions yields:

$$(1 - A)r^2 \frac{L_\odot}{4d^2} = \epsilon\sigma T^4 4\pi r^2 \iff T = \sqrt[4]{\frac{(1 - A)L_\odot}{16\epsilon\sigma d^2}}$$

The proxy for emissivity for asteroid infrared variation is 0.9 (V. Reddy *et al.* 2015). Albedos of asteroids can be readily obtained from the Small Body Database, see section 4. However, this information is not available for all bodies in the database. A common approach (see V. Reddy *et al.* 2015) is to use typical values for albedo from a given taxonomic class, when no information for a given asteroid is available. It has been applied in this project by creating a dictionary of average albedo values within each taxonomic class, where averages over all asteroids from a given class which had been present in databases under study, and for which a numerical value of albedo could be obtained from N. JetPropulsionLaboratory n.d.(a), were calculated.

One element remains unknown in the equation for temperature, namely the heliocentric distance d. From the Small Body Database, semi-major axis a and eccentricity e is obtained (section 4), and a could be a proxy for d, but in general $d \in [a(1-e), a(1+e)]$ and the exact value depends on the observation date and corresponding position of the asteroid on it's orbit. As will be discussed in 7.4, the ability to extract this information is somewhat limited, so whenever it is not available, a proxy of $d \approx a$ could be used, leading to a formula:

$$T = \sqrt[4]{\frac{(1-A)L_\odot}{16\epsilon\sigma a^2}}$$

in which all the numbers are known for a given asteroid of interest. To argument that such an approximation is quite harmless, consider the fact that most asteroids have $e \in [0, 0.3]$ (see e. g. Figure 11 right). Now, let's imagine an extreme case of $e = 0.3$ and a true equilibrium temperature of 200K (again, extreme case, for most asteroids it is somewhat lower). We could imagine two edge cases:

1. We used $d = a$ but in reality the asteroid was in aphelion, meaning that $d = a(1+e)$. In that case we would arrive at a temperature of $\sqrt{1+0.3} \cdot 200\text{K} \approx 228\text{K}$. The error of 28K would correspond to an $0.00075 \cdot 28 \approx 0.021$ error on BAR which would lead to about 1% error in the $\frac{\text{opx}}{\text{opx+ol}}$ calculation when using formulas introduced in section 8.

2. We used $d = a$ but in reality the asteroid was in peryhelion, meaning that $d = a(1-e)$. In that case we would arrive at a temperature of $\sqrt{1-0.3} \cdot 200\text{K} \approx 167\text{K}$. The error of 33K would correspond to an 0.025 error on BAR

Clearly even in such an exaggerated example the effect of not including a proper value of d in the calculation only distorts mineralogical content conclusions by about 1%, so they would not be severely misconstrued, and in an average case the discrepancy would be even lower. This is much less then the influence of temperature correction in general, not to mention the phase angle correction (see e. g. Table 5). In fact, such a high level of precision on temperature estimation is utopic anyways, since other approximations had been made in the way, such as assumption that an asteroid is a grey body with an emissivity of 0.9.

7.4 Observation-date-specific information

Apart from heliocentric distance d mentioned in 7.3, the phase angle α, although usually not too high as an observation would otherwise not be planned in the first place, depends on observation date. Determination of both of these observation-dependent parameters could be achieved by sending a relevant query to the ephemeris database Horizons (JetPropulsionLaboratory n.d.). Indeed, for the case of a phase angle, the appropriate link has been found and implemented in the project, presented below with parts which need to be altered from one case to another in bold:

https://ssd.jpl.nasa.gov/api/horizons.api?format=text&COMMAND="**Name=Itokawa**"&OBJ_DATA="NO"&
MAKE_EPHEM="YES"&EPHEM_TYPE="OBSERVER"&CENTER="geo"&TLIST="**2001-03-28**"&QUANTITIES="24"

where :

- the beginning of the link sygnalizes a request to the Horizons API
- *format=text* informs that the ephemeris information should be provided in a simple text format, which can easily be parsed by the code
- *COMMAND=* sets the conditions of the search. There are two options here:
  - Putting the designation number of the asteroid after this keyword appears to give the desired effect for most numbers higher than 10. For example, replacing **"Name=Itokawa"** with **25143** in the link above would generate the same result. However:

        * Ephemeris for asteroids 1 Ceres, 2 Pallas, ..., 10 Hygiea cannot be found in this way because the numerical code gets mistaken for a numbering referencing one of the big bodies of the Solar System.
        * Referencing the designation number can lead to ambiguities in general, because the Horizons database uses it's own set of ID numbers to reference some relevant points in the Solar System. For example:
            · An asteroid 9 Metis, which as was mentioned above cannot be found by calling number 9, can be found under a number of 516, instead of 516 Amherstia (a fact discovered accidentally while studying Fornasier spectra (9.1).
            · An attempt to look up ephemeris for asteroid 399 Persephone by referencing it's number would be futile, because the Horizons database references the Earth with a number 399 and would respond with an error message *Observer table for observer=target disallowed* to such a request.
    – A far more error-proof alternative is to reference asteroids by name, e. g. *COMMAND="Name=Vesta"* to enquire about 4 Vesta. Problems with the first option motivated a choice that the function in the code Morawski 2023-2024 would choose to reference the asteroid by the name, unless it is not known (has not been extracted from the Small Body Database (N. JetPropulsionLaboratory n.d.(a) because the asteroid is nameless)) - only asteroids without a name would be referred to by their number. This is a satisfyingly safe strategy because asteroids without names have high designation numbers, which should not have a second meaning in the Horizons database. However, it later turned out that there are rare cases of names which lead to ambiguity: name Lee can refer to either 3155 Lee or to a comet C/1999 H1. To resolve these rare cases, if the code fails to obtain information using the name, the function recursively calls itself with a flag telling it to use the number approach instead.

- *OBJ_DATA="NO"* removes an unnecessary part of the output corresponding to information about the object
- *MAKE_EPHEM="YES"* requests generating an ephemeris
- *TYPE="OBSERVER"* specifies the type of ephemeric (observables)
- *CENTER="geo"* places a hypothetical observer in the center of the Earth
- *TLIST=* precedes a list of dates for which we want to obtain an ephemeris. The single date of observation in a yyyy-mm-dd format should be provided.
- *QUANTITIES="24"* specifies the type of information we want to obtain, the numerical code 24 referencing a Sun-target-observer angle, i. e. phase angle α

Getting the heliocentric distance for higher accuracy in the temperature correction formula (7.3) is a bit more complicated, because there is no single *QUANTITIES* reference number that would output this information. It is necessary to change the link in several ways and get:

https://ssd.jpl.nasa.gov/api/horizons.api?format=text&COMMAND=**"Name=Itokawa"**&OBJ_DATA="NO"&
MAKE_EPHEM="YES"&EPHEM_TYPE="VECTORS"&CENTER="@Sun"&TLIST="**2001-03-28**"&QUANTITIES="1"

where :

- the first part is same as before
- *TYPE="VECTORS"* specifies the type of ephemeric as vectors, meaning that cartesian coordinates of the object in 3D space will be returned
- *CENTER="@Sun"* places the origin of the coordinate system in the center of the Sun
- *TLIST=* same as before
- *QUANTITIES="1"* return the X, Y, Z coordinates in kilometers when *TYPE="VECTORS"* is being used

As we can see, the desired value of d is not returned directly, but can be computed as $d[m] = 1000\sqrt{X^2 + Y^2 + Z^2}$.

Automatic generation of such links, and segmenting the output string to isolate the single numerical value of the phase angle α (in degrees) and the values of X, Y, Z for computing d, respectively, has been tested with success on the example of 25143 Itokawa and on other examples from the IRTF database. Ergo, the investigation regarding the ability of accessing relevant observation-date-specific information in a programmatic way has been concluded with a positive answer, since the Horizons API (JetPropulsionLaboratory n.d.) does make it possible. However, the essential

piece of information in order to be able to make such queries are observation dates, which cannot be associated with each spectrum found at PlanetaryDataSystem n.d. in the most general case.

Indeed, data files with spectra do not contain information about the observation date and conditions. In many cases it needs to be accessed separately, usually from a table in a relevant publication. For small surveys the task is simple enough, for example relevant information for the Fornasier dataset can be found in Table 1 of Fornasier *et al.* 2010. While not entirely an automated process, it would not take too much work to put this information into a text file and translate it into a dictionary of asteroid - observation date pairs, which could then be used by the code to send relevant queries about date-specific information to the Horizons database. In this project, it was only done for one example of 516 Amherstia (9.1.2). For larger datasets, however, the task would be daunting in the very least and defies the purpose of automating the work which is the main drive of the entire project.

Luckily, in the IRTF database (2.5) a very straightforward solution to keeping this highly relevant information available has been found by providing it in the filename of each spectrum (the fact that numbers in the filenames indeed corresponded to observation dates has been verified for a few cases from Moskovitz *et al.* 2010). This should serve as an example for any current and future surveys, since it is pivotal for the task of gathering wisdom on mineralogical contents of thousands of asteroids that not only would high-quality spectra be available, but relevant information to make the corrections necessary (i. e. observation date) could be accessed with ease. Additionally, it would be advisable for surveys designed with the objective of serving such goal to record spectra during oposition, so that the necessity for phase reddening correction could be reduced as much as possible - even if the correction is being made, it will always be a source of error in itself, especially with the current state of knowledge on the matter with only one insufficiently tested correction formula available (see 7.2).

8 Formulas for mineralogical content

There exists a number of empirical formulas (created based on analysis of meteorite spectra), which try to infer percentages of olivine and pyroxene content based on band center positions and band area ratio. Gaffey *et al.* 1993 defines an area within the BI - BAR parameter space where S(IV) meteorite parent bodies should be found, and provides an algorithm for calculating wollastonite and ferrosilite percentage content among the pyroxenes, assuming that the absorption features are due to pyroxene only (and not olivine).

A more general formula has been proposed by Cloutis *et al.* 1986, which is supposed to compute ratios between olivines and orthopyroxenes within a mixture of the two, based on band area ratio alone:

$$\frac{\text{opx}}{\text{opx}+\text{ol}} = 0.417 \cdot \text{BAR} + 0.052 \ , \ \frac{\text{ol}}{\text{opx}+\text{ol}} = -0.417 \cdot \text{BAR} + 0.948$$

where opx an ol refer to orthopyroxene and olivine content, respectively. The formula is supposed to work well in the range of $\frac{\text{opx}}{\text{opx}+\text{ol}} \in [0.1, 0.9]$. However, objections have been made that the formula only applied well to mixtures containing a single pyroxene, which is not the case for most asteroids.

A different formula, which can be relevant for more asteroids, has been proposed by Dunn *et al.* 2010, who also developed formulas to estimate fayalite and ferrosilite percentage content among olivines and pyroxenes respectively, assuming that the body under study is an ordinary chondrite. The full set of Dunn *et al.* 2010 formulas are:

$$\frac{\text{opx}}{\text{opx}+\text{ol}} = 0.242 \cdot \text{BAR} + 0.272 \ , \ \frac{\text{ol}}{\text{opx}+\text{ol}} = -0.242 \cdot \text{BAR} + 0.728$$

$$\frac{\text{fa}}{\text{ol}} = -12.849 \cdot (\text{BI})^2 + 26.565 \cdot \text{BI} - 13.423$$

$$\frac{\text{fs}}{\text{px}} = -8.791 \cdot (\text{BI})^2 + 18.249 \cdot \text{BI} - 9.217$$

A formula for the high calcium (wollastonite-like) pyroxene content in correlation with band II center has also been sought after, but a fit found has not been satisfactory and the idea of a trustworthy correlation has been dismissed (low R^2 value of the fit).

The Dunn *et al.* 2010 formula has one oddity about it, which should be kept in mind. Logically possible values of $\frac{\text{opx}}{\text{opx}+\text{ol}}$ would be between 0 and 1, which would correspond to BAR$\in [-1.124, 3.008]$. On the other hand, BAR

should represent a ratio of two areas, so negative BAR (which would be necessary to get $\frac{\text{opx}}{\text{opx+ol}} < 24.2\%$ according to the Dunn *et al.* 2010 formula) does not make sense. Clearly the test sample for which the formula has been fit did not include meteorites with very low orthopyroxene content. Indeed, looking into the Dunn *et al.* 2010 paper one can confirm that all meteorites from the sample had $\frac{\text{opx}}{\text{opx+ol}} \in [45\%, 70\%]$. While the linear fit has been quite convincing ($R^2 = 0.73$), there is no evidence whatsoever that it could be extrapolated. On the contrary, it feels as a logical necessity that inverting the formula should be yielding nonnegative band area ratios for all $\frac{\text{opx}}{\text{opx+ol}} \in [0, 1]$, which is not the case for an extrapolation in the lower range.

One seemingly reasonable approach to tackle this dilemma would be to extrapolate it in a linear fashion between $\frac{\text{opx}}{\text{opx+ol}}(BAR = 0) = 0\%$ and, say $\frac{\text{opx}}{\text{opx+ol}}(BAR = 0.75) = 45.35\%$. However:

- It is generally naive and nonscientific to extrapolate any formula to a range where no datapoints exist, unless other arguments can support such a claim (for example extrapolating a function with established periodic temporal behavior into the future)
- Even if a guess $\frac{\text{opx}}{\text{opx+ol}}(BAR = 0) = 0\%$ would be reasonable (lack of pyroxenes would result in the absence of 2μm band, an asteroids with olivines only would have solely the $\sim 1\mu$m feature), and the other end of the $[0\%, 45\%]$ interval can be supported by the Dunn *et al.* 2010 study, we cannot be sure if the relashionship is anything close to linear in that range. Maybe the Dunn *et al.* 2010 formula works as long as it corresponds to nonnegative BAR $\left(\frac{\text{opx}}{\text{opx+ol}} > 24.2\%\right)$, and the rest of the interval corresponds to 0 BAR, meaning that the $\sim 2\mu$m band would not manifest itself in low-pyroxene asteroids? That does not sound very likely, but maybe the relationship does extrapolate up to some point, and only then changing the slope to match $\frac{\text{opx}}{\text{opx+ol}}(BAR = 0) = 0\%$ would be adequate, but if, where is that point?
- Things get even more fuzzy when we consider the fact that applying temperature and phase angle corrections (7.2) to parameters inferred from asteroid spectra can sometimes result in negative BARs, as we will see for two Sa class asteroids analyzed in **??**. Presumably, two erroneous effects come into play here:
  - Corrections introduced in 7.2 are flawed, in particular the phase angle correction is supported by a study of one asteroid only in a limited range of phase angles, and probably the correction is overestimated in a general case, see discussion in 7.2.
  - Non-scientific extrapolation of the Dunn *et al.* 2010 formula

  Ironically, both of them work in "the same direction", i. e. pushing BAR towards negative values, and have a potential of cancelling each other out - orthopyroxene contents derived from putting negative BARs into the Dunn formula (as miscalculated with the exaggerated phase angle correction from Vishnu Reddy *et al.* 2012), maybe be closer to actual values when influenced by both calculation erros than if only one of them were in the play.

Taking into account all of this uncertainty regarding how to infer $\frac{\text{opx}}{\text{opx+ol}}$ when BAR< 0.75, this project applies the formula in it's original design in every case (even if temperature and phase angle corrected BAR is negative), but gives less credibility to mineralogical contents which do not fit in the range corresponding to BAR$\in (0.75, 1.75)$, which is roughly the range of the Dunn *et al.* 2010 sample.

A different set of formulas has been found by Burbine, Buchanan, and R. P. Binzel 2007 for howardites and eucrites and is advised to be used when analyzing V class asteroid spectra. On the downside, no correlation between BAR and $\frac{\text{opx}}{\text{opx+ol}}$ ratio has been found. On the other hand though, both ferrosilite and wollastonite contents can be computed, allowing a placement of such an asteroid on the trenary diagram of pyroxenes. Moreover, two sets of formulas, dependent on each of the bands independently, have been proposed, providing a ready tool to verify results by cross-checking the two.

Formulas based on the first band are:

$$\frac{\text{fs}}{\text{px}} = 10.234 \cdot \text{BI} - 9.1382$$

$$\frac{\text{wo}}{\text{px}} = 3.961 \cdot \text{BI} - 3.6055$$

Formulas based on the second band are:

$$\frac{\text{fs}}{\text{px}} = 2.0586 \cdot \text{BII} - 3.643$$

$$\frac{\text{wo}}{\text{px}} = 0.79905 \cdot \text{BI} - 1.483$$

9 Absorption feature search and mineralogical analysis for the data

9.1 Absorption bands in the spectra of M type asteroids from the Fornasier database

9.1.1 Searching for $\sim 0.43\mu$m and $\sim 0.9\mu$m absorption bands

As mentioned in the section 2.1, one of the goals of the Fornasier survey was to look for absorption features in the spectra of metallic asteroids, specifically for a feature around 0.9μm attributed to low iron otrthopyroxenes (the pyroxene $\sim 1\mu$m wavelength is known to reside in the lower part of it's typical wavelength range, i. e. closer to 0.9μm, if the iron content is low, especially for low calcium pyroxenes, see Adams 1974) and a $\sim 0.43\mu$m feature of disputable origin (various minerals had been suggested as responsible).

Taxonomic class	Total number of asteroids	$\sim 0.43\mu$m band detection cases		$\sim 0.9\mu$m band detection cases		Both bands detection cases	
		$S = 29$	$S = 11$	$S = 29$	$S = 11$	$S = 29$	$S = 11$
Sq	1	1	1	1	1	1	1
X	1	-	-	-	-	-	-
Xe	2	-	-	-	-	-	-
Xk	2	-	-	1	1	-	-
T	12	4	4	8	6	3	2
D	8	4	1	-	1	-	-
L	4	2	1	-	-	-	-

Table 4: Cases of absorption band detections in the spectra of M type asteroids from the Fornasier database (2.1) divided by taxonomic classes. Here assuming that 125 Liberatrix belongs to class Xk (see discussion in 5.2,5.3). Number of detections is provided for each band o interest for each choice of a smoothing window size S considered (see 6.1). Dashes indicate no detections. We can conclude that within the spectra in the Fornasier database the bands are most likely to be found among the classes Sq (100% of cases, but this refers to one asteroid only) and T (75% of cases with at least one of the bands found for $S = 29$). Additionally, it is apparent that a bigger choice of the averaging window size $S = 29$ increases the likelihood of absorption band detection.

An algorithm from 6.2 was run over the database to search for these two absorption bands, trying out both a small ($S = 11$) and big ($S = 29$) choice of the smoothing window size. Results are summarized in Table 4. A closer look at the plots suggests that in fact only the cases corresponding to taxonomic classes Sq and T can be considered accurate, scarce detections in other classes are rather due to the fact that a search for these subtle absorption bands requires setting lower thresholds on parameters of minimum width and depth, leading to a higher probability of a false positive. Some curiosities could be noted, for example a fact that among the 2 Xk asteroids from the database, only 125 Liberatrix has a detection of a 0.9μm band when using $S = 11$, but when using $S = 29$, it was only the other one, 161 Athor. However, such conclusions are ultimately meaningless, because a subtle change of the thresholds on the width and depth of absorption bands would dismiss those vague cases, their lack of validity being exemplified on Figure 17. It is not the case for classes Sq and T, on which we can now focus.

An in-depth look into the T class representatives with absorption band detections has confirmed an observation from Table 4 that a choice of $S = 29$ as the smoothing window for the analysis of this dataset is superior. Compared to $S = 11$, the window of $S = 29$:

- Has more detections in the taxonomic class T (see Table 4)
- The spectra with detections of the first absorption band near $\sim 0.43\mu$m, if they retain that feature for the larger smoothing window, form a more robust case for the presence of that band (Figure 18), as the dimensions of the feature are no longer comparable with the noise (which has been reduced in the smoothing process)
- Leads to the same conclusions regarding the position of the center of the $\sim 0.9\mu$m band, with the sole exception of 216 Kleopatra, where it is overestimated (lower part of Figure 19)
- Tends to overestimate the areas of the features (see Figure 19), but these are not relevant for the analysis in this section anyway.

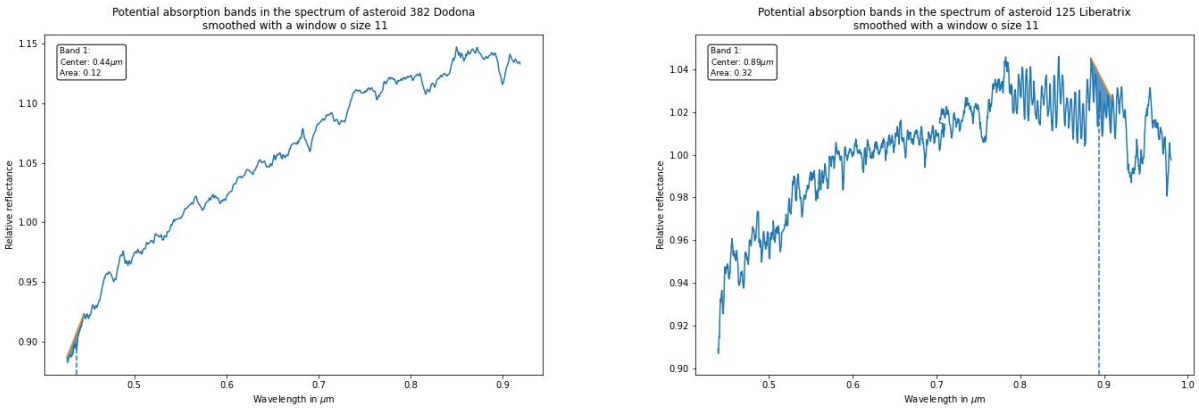

Figure 17: Examples of inaccurate absorption band detections for objects outside of taxonomic classes Sq and T in the Fornasier database. **Left:** L class asteroid 382 Dodona has a band detected at 0.44μm. In reality, the small width and shallow depth of the band make it extremely unlikely that the band is genuine, especially considering that the noise level is high. **Right:** a band at 0.89μm is found in a spectrum of an Xk object 125 Liberatrix. In reality, it is not even the most likely candidate for a band in that region of the spectrum, when analyzed by a human. Moreover, among the three similar spectra of this object found in the Fornasier database, only one of them has the feature detected, and only for the smoothing window of size 11. Hence, both cases should be disregarded as false positives.

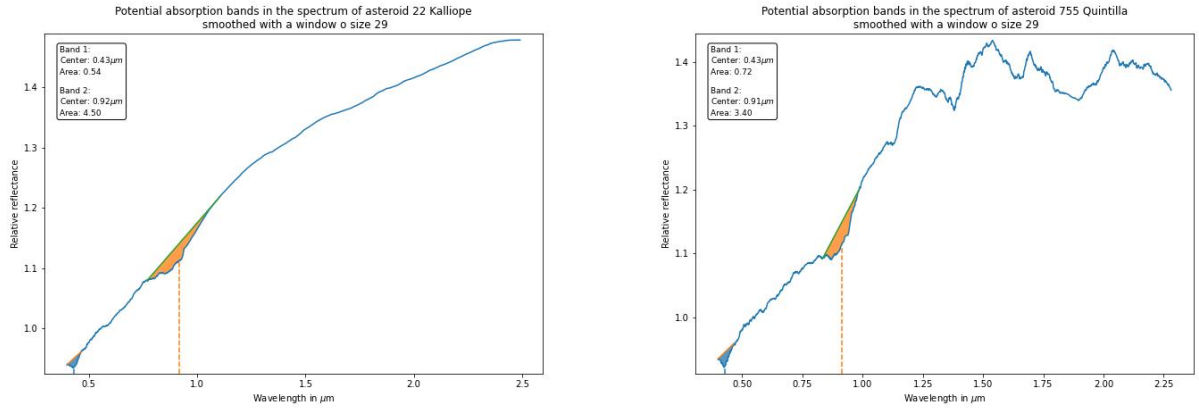

Figure 18: Two most believable detections of the 0.43μm absorption band among the spectra from the Fornasier database. Both 22 Kalliope and 755 Quintilla retain a small, but clearly defined dip at this exact wavelength even after applying average smoothing with a window of size 29.

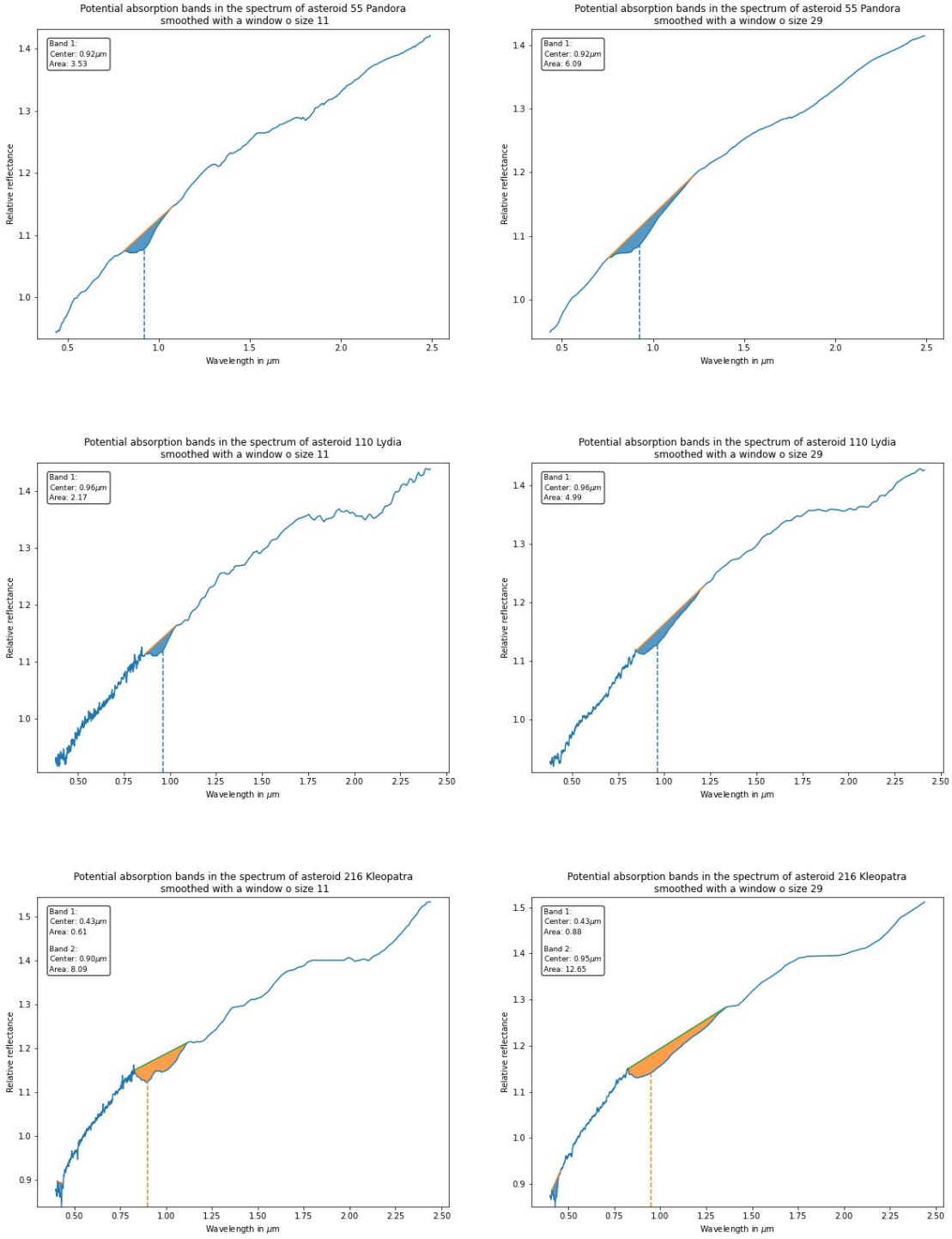

Figure 19: Comparison between the positions and dimensions of absorption features detected in spectra of three T class asteroids from the Fornasier database. Spectra on the left are results of average smoothing (see 6.1) of original spectra with a window of size 11, for the right side it is 29. The bigger window choice leads to a smoother spectrum, but areas of the 0.9μm absorption bands seem to be somewhat overestimated. A more basic property of central wavelength, however, is the same for both choices of smoothing windows, except for 216 Kleopatra, where it is shifted towards longer wavelengths (but it is the only such case among all detections for T class asteroids from this database).

9.1.2 Olivine/pyroxene content analysis for 516 Amherstia

The sole Sq class representative 516 Amherstia, features both absorption bands, and it stands out from the cases discussed above with a much bigger depth, and consequently area, of the $\sim 0.9\mu$m feature (here $w_{\text{cen}} = 0.93\mu$m). From the taxonomic point of view, it is only to be expected, since S-complex spectra are known for their features, whereas T class spectra are generally relatively featureless. The $\sim 0.43\mu$m (here 0.44μm) is not as clearly defined as for the two T class cases illustrated on Figure 18, so it might be a false positive.

The fact that 516 Amherstia belongs to S-complex,as well as the fact that the spectrum showcased in the Fornasier database extends out to the infrared, have inspired a search for the $\sim 2\mu$m pyroxene band as well. As Figure 20 reveals, while a 0.44μm band is questionable in the very least, a very clearly defined 1.99μm absorption band is present. Additionally, a comparison between two smoothing window sizes is made again, leading to the opposite conclusion then in the previous section: for bigger absorption features and for studies where accurate estimates of band areas are desired, a choice of $S = 11$, which preserves the original shape of the spectrum better, is desired.

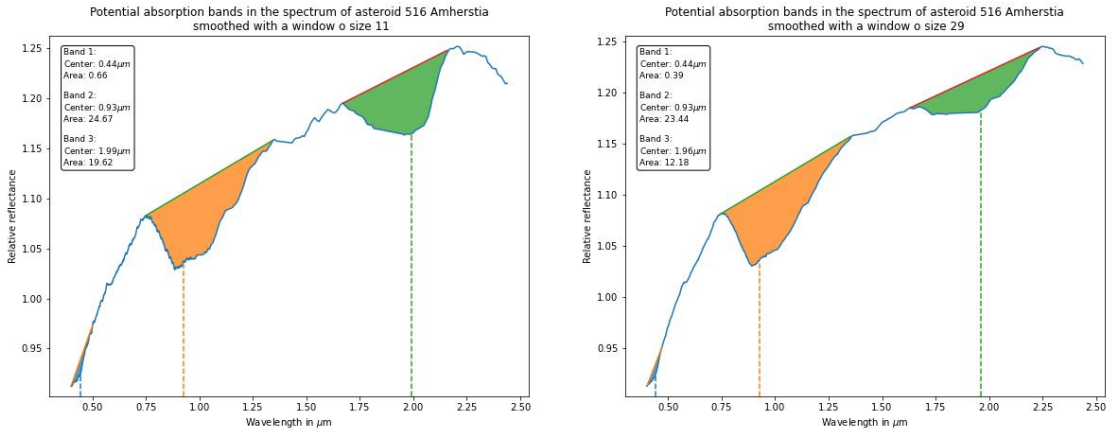

Figure 20: Absorption bands found by the algorithm (6.2) in the spectrum of an Sq class asteroid 516 Amherstia from the Fornasier database, using (**left**) a smoothing window of size $S = 11$ and (**right**) a smoothing window of size $S = 29$. Three bands were searched for in this particular case. The 0.44μm band is not very convincing and might be a false positive. The 0.93μm is a prominent feature, but so is the 1.99μm feature when using $S = 11$. The alternative choice of $S = 29$ seems to distort this last band quite severely and reduce it's area, which makes the results from applying mineralogical formulas from 8 more beleviable for the parameters found on the left plot.

For an S-complex asteroid, Dunn *et al.* 2010 formulas seem to be a better choice of mineralogical content formulas to use, among those introduced in section 8. In fact, inferred from the $S = 11$ derived parameters from Figure 20 left, band I center at 0.93μm and BAR= 0.795, make it the clearest case of S(IV) asteroid among all covered in this study, according to the region defined by Gaffey *et al.* 1993. Bear in mind that bands I and II as spoken of in the mineralogical study context (sections 7,8) are bands 2 and 3 on Figure 20, because band 1 refers to a potential 0.44μm band there.

Results of applying Dunn *et al.* 2010 formulas are presented in Table 5. For completeness, calculations were made with both choices of the smoothing window, and we can confirm that the decision influences a resulting $\frac{\text{opx}}{\text{opx+ol}}$ ratio quite significantly. Although a choice of a bigger smoothing window underestimates the orthopyroxene content, not taking corrections mentioned in section 7 leads to an overestimation of a similar extent. As discussed in section 7, making the phase angle correction for a general case is difficult in the project as it is. However, observation dates for the Fornasier dataset can all be found in one place, namely Fornasier *et al.* 2010, so we can verify that the spectrum of 516 Amherstia was obtained on the 19^{th} of November 2004, and a corresponding phase angle of 11.9581° can be obtained with the methodology presented in 7.4.

Calculation method	$\frac{\text{opx}}{\text{opx+ol}}$	$\frac{\text{fa}}{\text{ol}}$	$\frac{\text{fs}}{\text{px}}$
$S = 29$, no correction	40.0%	16.1%	14.6%
$S = 29$, temperature correction	37.2%	18.1%	16.0%
$S = 29$, temperature and phase angle correction	28.8%	18.1%	16.0%
$S = 11$, no correction	46.5%	16.2%	14.6%
$S = 11$, temperature correction	44%	18.2%	16.0%
$S = 11$, temperature and phase angle correction	**35.4**%	**18.2**%	**16.0**%

Table 5: Mineralogical context calculated with equations from Dunn *et al.* 2010 (see section 8) for the spectrum of 516 Amherstia from the Fornasier database (2.1). Different ways of evaluating input parameters for the formulas are compared: two choices of smoothing window (6.1), temperature and correction as described in section 7. The bottom-most line should be considered the most reliable. We can see that the bigger smoothing window choice tendentiously underestimated orthopyroxene content $\frac{\text{opx}}{\text{opx+ol}}$, but did not have an effect on the fayalite and ferrosilite abundancies (because they depend on the first band only, which is found at 0.93μm in both cases). We can also see that the temperature correction matters, as it can change the percenteges by a few, but even more so is the phase angle correction relevant.

9.2 Olivine/pyroxene bands in the IRTF spectra

9.2.1 Searching for the bands and choosing the smoothing window size

As mentioned in 2.5, one of the main reasons for adding the IRTF database to the project has been an interest in looking for olivine and pyroxene features in a larger set of spectra which do display them, which required finding a database which covers the range of the second pyroxene band ($\sim 2\mu$m). The very case of detecting the two bands characteristic of a mixture of the two minerals, i. e. $\sim 1\mu$m and $\sim 2\mu$m, in the IRTF data representing the taxonomic class V known for the most prominent absorption bands (see 5.1), has been used for designing and evaluating the algorithm (6.2, e. g. Figure 15).

Running the absorption band algorithm on the spectra from the IRTF database with the default parameters corresponding to these two bands of interest has shown (in accordance with expectations based on the theory of taxonomic classification, see 5.1) that spectra from the end member taxonomic classes V and R, as well as most S-complex classes (S, Sa, Sr, Sv) exhibit these two absorption features, and they can be reliably detected using the algorithm described in 6.2. Indeed, both bands were found for a vast majority of asteroids from these six taxonomic classes (83% with the average smoothing filter (6.1) of $S = 29$, 75% with $S = 11$), and at least one of the two bands has been found in the remaining ones, except for a total of two cases for which had no absorption band detections for $S = 29$ (and only a very questionable detection of the first band when using the smaller window of $S = 11$): nameless S class near Earth object 29075 and an Sa class main belt asteroid 322 Phaeo, the spectra of both of which also do not hint at plausible absorption bands upon human inspection of the plots.

For the actual analysis, initially, both choices of smoothing windows needed to be considered, for in this section we will also be interested in measuring band areas, and as we have seen in section 9.1 the choice of $S = 29$ can lead to an overestimation of those (Figure 19). However, since the $\sim 1\mu$m and $\sim 2\mu$m olivine/pyroxene features under analysis here are generally much deeper, broader features than the once investigated in 9.1, the effect seems to dissipate. This seems to be an argument in favor of choosing the larger smoothing window, especially since the contrary choice seems to sometimes yield dissatisfactory results of non-physical detections (e. g. Figure 17). And yet, as exemplified by Figure 21, there is another problem appearing in this case, where the choice of a bigger window actually tendentiously leads to slight underestimation of band areas in a non-uniform manner. The underestimation of the $\sim 1\mu$m band area is stronger because that band occupies an area close to the end of the available data, so frequently corresponding $w_{\min}$ is equal to w_1 (in the sense of symbology from section 6.1, i. e., it is the bluemost datapoint), and as a result it becomes averaged together with more points of lower reflectance, without anything on the other side to balance out this average; moreover the slope of the spectra is generally higher in that area, so the differences between the values within an averaging window, and consequent attenuation of extreme values, is stronger as well. This underestimation of the Band I area would distort calculations of the band area ratio, which is relevant for mineralogical analysis, so a choice of $S = 11$ will be favored in this section. A more general conclusion is that a choice of $S = 29$ always tends to underestimate band areas (a similar problem occured for 516 Amherstia in 9.1.2, see Figure 20, except there it was mostly the second band the area of which has become underestimated). Although there are slightly more omission and comission cases when running the absorption band detection algorithm (6.2) with $S = 11$ as opposed to the larger window $S = 29$, the erraneous results for band areas would render calculations with the latter inaccurate. Therefore,

for olivine/pyroxene absorption the smaller window should be chosen, even if it may exclude a few asteroids from the analysis. Examples of spectra representing several of the taxonomic classes of interest, as analyzed with the algorithm with a smoothing window $S = 11$, is presented on Figure 22.

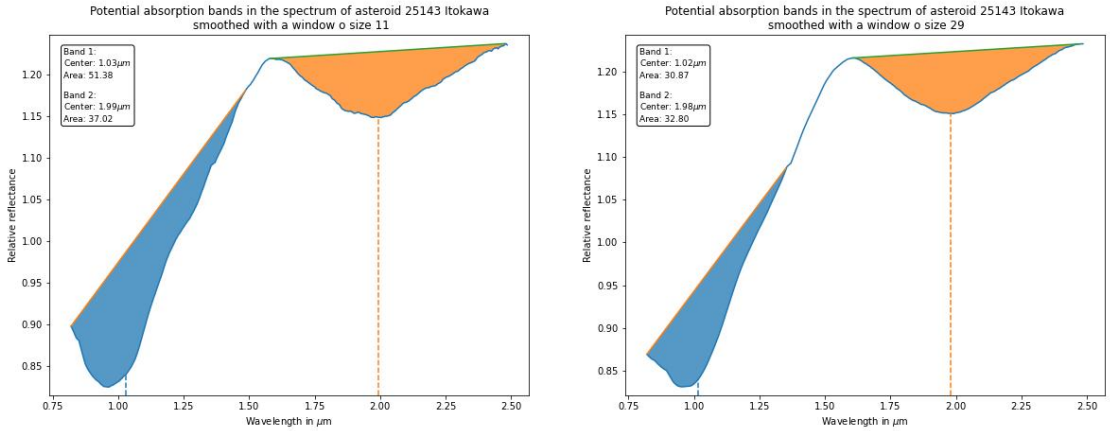

Figure 21: Problem with the bigger smoothing window choice when searching for olivine/pyroxene absorption bands, exemplified by the case of an Sa class asteroid Itokawa spectrum from the IRTF database. Areas under the bands end up being underestimated for the smoothing window coice $S = 29$, with the first band (1.03μm) being affected more severely than the second one (1.99μm): the area of the second band as measured with $S = 29$ is reduced by $\frac{37.02-32.8}{37.02} \approx 11\%$, whereas the area of the first one by $\frac{51.38-30.87}{51.38} \approx 40\%$. Such disproportionate treatment of band areas would lead to severe distortion of the band area ratio calculation (in this case it would be almost 1.5 times higher when calculated using $S = 29$ compared to $S = 11$). Even though the spectra smoothened with $S = 11$ retain a some residual noise, overall, we can consider the calculations regarding the absorption bands, particularly band areas, as more believable than for $S = 29$ (clearly the band identified on the left plot looks like a better approximation of the feature), so the smaller smoothing window choice will be favored for the analysis in section 9.2.

9.2.2 Band centers and band area ratio

Olivine/pyroxene band position information is commonly used to place asteroids and meteorites on the BI - BAR plot, where the position of the first band (BI) is plotted against the band area ratio BAR. For the 6 taxonomic classes of interest in the IRTF dataset, such a plot is presented on Figure 23, showing also how the temperature and phase angle corrections described in section 7 affect positions of points in that parameter space. Although the temperature correction is important for mineralogical analysis, as we have already seen in 9.1.2, it's impact on the general appearance of the BI - BAR - BII parameter space is very slight. The effect of phase angle corrections is starker, and suspicious negative BAR values arise in several cases. The two Sa class objects present in the dataset stand out with exceptionally long band center wavelengths. Quite a huge scatter is visible among other data points, especially for the V class objects, but the band I positions seem to be rather consistent within each taxonomic class. The latter observation is confirmed by analysis on Figure 24, where whisker box plots of band I and band II position ranges for each class of interest (excluding class S where the $\sim 1\mu$m band was not detected in any of the asteroids). The former inspired to generate plots shown on Figure 25, where a look is taken into whether any correlations between asteroid intrinsic physical properties such as diameter and rotational period and the band area ratio could be found, with a conclusion that no strong trends can be inferred, but a subtle hint for a positive correlation in both cases is present.

Having made observations about the nature of the absorption features in terms of their positions and areas, it is time to attempt linking these results with mineralogical composition of asteroids, which will be covered in the following sections.

9.2.3 Mineralogical analysis of S class asteroids

First, let's take a look at S-complex objects. A claim that using Dunn *et al.* 2010 mineralogical formulas is justifiable for all of those is rather far-fetched, since they do not necessarily all bear similarity to ordinary chondrite meteorites, and band area ratios of most them fell outside of the range corresponding to $\frac{\text{opx}}{\text{opx+ol}} \in (45\%, 70\%)$ of the Dunn *et al.*

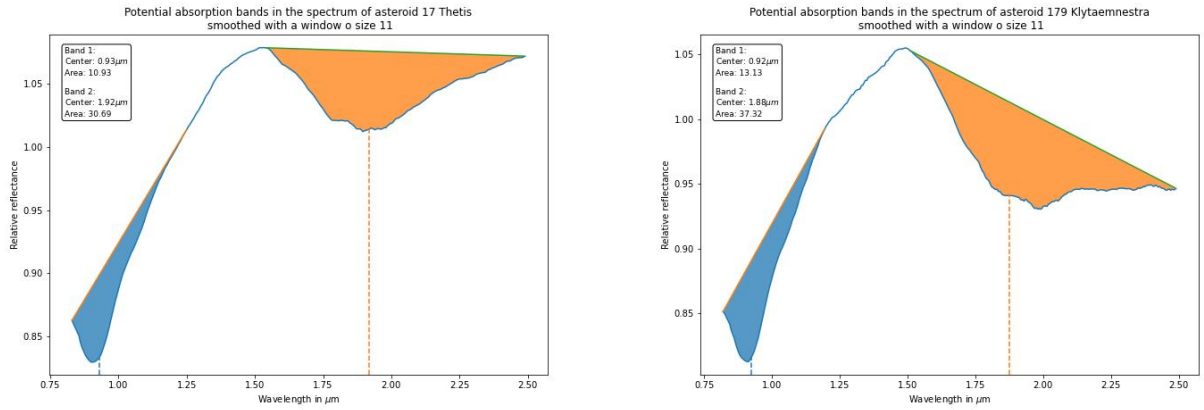

Figure 22: Examples of IRTF spectra with olivine/pyroxene absorption bands as detected by the algorithm (6.2. **Left:** Sr class asteroid 17 Thetis. The class Sr is supposed to be a bridge between classes S and R, and typical R class spectra look quite similar, except the second band region have a bit more of a negative slope. 17 Thetis is the only Sr class asteroid in the IRTF dataset (see Figure 10). **Right:** Sv class asteroid 179 Klytaemnestra. Sv asteroid spectra have a steeper downward slope in the area of the second band.

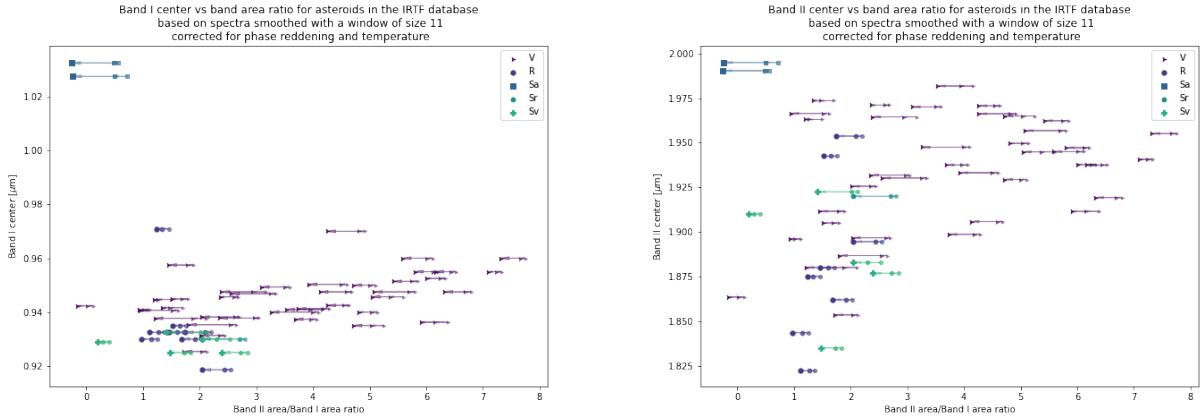

Figure 23: **Left:** Band I center vs band area ratio for asteroids from the IRTF database of several taxonomic classes which display olivine/pyroxene absorption classes. **Right:** Same but with band II on the vertical axis. Each triplet of points corresponds to a single asteroid with arrows portraying how the values change when taking into account first the temperature correction only, then phase angle correction as well. Whenever multiple spectra were available for the same asteroid, band center positions and band area ratios were averaged over all spectra to calculate the position on the plot. Spectra in which only one or no absorption bands were detected had been excluded from the calculation, which is a reason for absence of class S representatives, for which band I was not detected in any case when using a smoothing window $S = 11$. Band I positions are quite similar within each taxonomic class, whereas band II positions exhibit larger spread, as will be better shown on Figure 24. The effect of the temperature correction for the overall structure of the plot is subtle. Some of the points move slightly to the left, as the temperature-corrected band area ratios become lower. This is bound to relate to asteroids on larger orbits, the temperature of which would be lower. Even though Vishnu Reddy *et al.* 2012 has been used for the temperature correction in the case of V and R classes, so band I position shifts could be possible, and all temperature corrections affect band II position, they turn out to be so tiny that the changes in ordinate coordinate eludes human eye comparison on both plots. Phase angle corrections are generally larger and as discussed in 7.2 they might be overestimated. In particular, in three cases phase angle corrections push the points to the paradoxical region of BAR< 0 (two Sa class asteroids 25143 Itokawa and 43 Ariadne, which will be discussed in **??** and a V class asteroid 9481 Menchu).

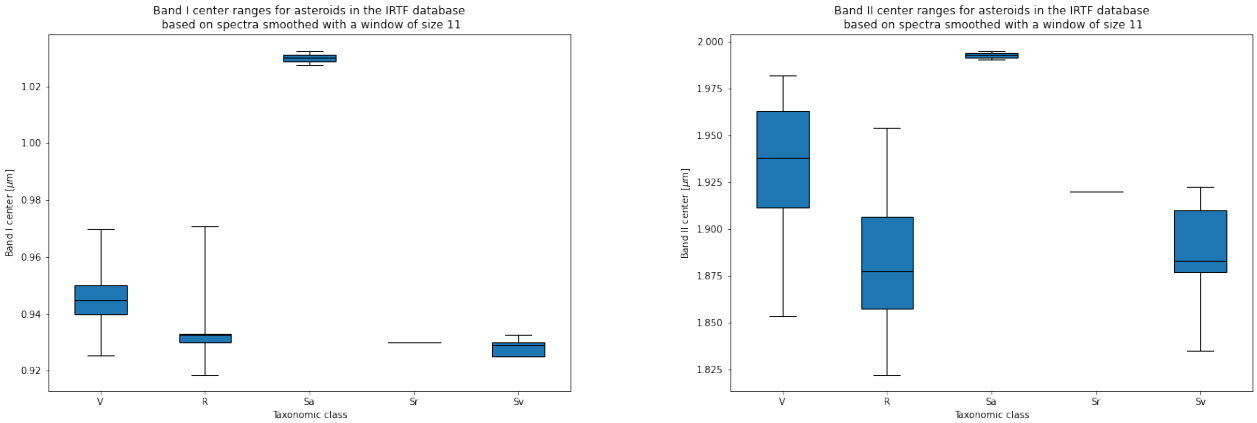

Figure 24: Whisker box plots for band I (**left**) and band II (**right**) positions for the five taxonomic classes for which both bands were consistently detected in the IRTF data. Blue boxes comprise 50% of the data centered around median band center for each class and the lines extend up and down to most extreme values. For the Sr class, a single horizontal line is drawn as there is only one representative (17 Thetis, see Figure 22 upper right). The Sa class has only two representatives, hence the median is actually a mean of two values, which correspond to the ends of the whisker lines. Nevertheless, the small width of that interval goes to show that both band I and band II positions are very consistent between the two specimen (43 Ariadne and 25143 Itokawa). As for the other three classes, positions of the first band are quite consistent within the dataset, especially for the Sv class, but the positions of the second band span much bigger ranges (notice that not only are the boxes larger on the right plot, but also the range of the ordinate axis is almost twice as big there). Temperature correction from section 7 is taken into account here, but as pointed out at Figure 23, it's impact on band positions is actually negligible.

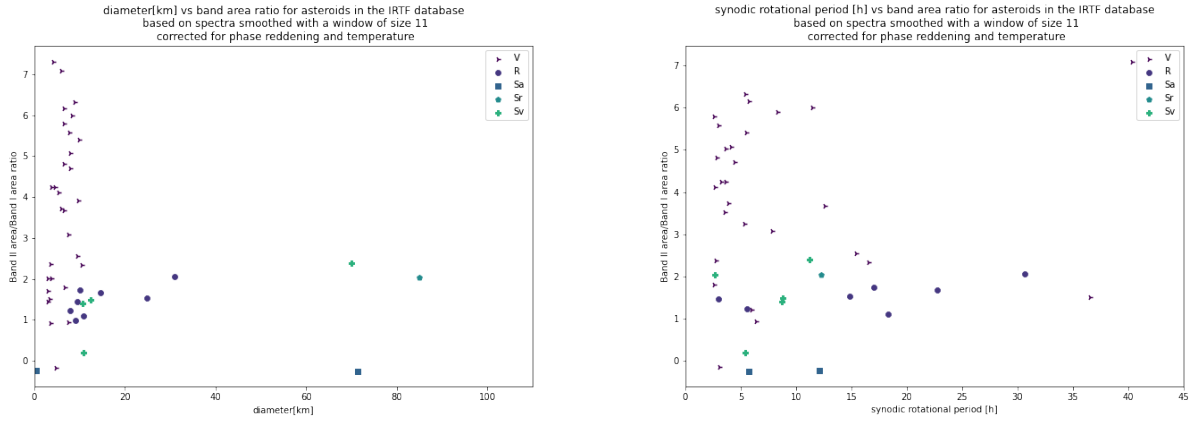

Figure 25: Temperature and phase angle corrected and area ratio plotted against asteroid diameter (**left**) and synodc rotation period (**right**). Abscissa data was extracted from the Small Body Database (N. JetPropulsionLaboratory n.d.(a)) with methodology described in section 4. Extreme outliers are excluded from the plot for clarity (4 Vesta with it's huge diameter of 525.4km in the case of the left plot and 2045 Peking with a rotation period of 158.7 hours on the right plot, both V class asteroids). The number of data points is insufficient to make any firm conclusions, but subtle positive correlations between the band area ratio and both diameter and rotational period of the asteroid might be the case for the R and Sv classes. Behavior of the band ratio for the V class seems to be rather chaotic, which is probably due to the fact that a variety of minerals are expected to occur in V class objects (see 5.1) which may contribute to both absorption features in ways which differ from one case to another, in particular no mineralogical content estimation formulas exist which would involve band area ratio for V class asteroids (see section 8).

2010 sample (see discussion in section 8). Nevertheless, it seems more reasonable to treat S-complex objects as ordinary chondrite proxies then as potential howardite-eucrite-like bodies (which is only supported for V class objects which will be analyzed in section 9.2.4), so the Dunn *et al.* 2010 formula is the most adequate available. Results of applying the formula are presented in Table 6.

Asteroid	$\frac{\text{opx}}{\text{opx+ol}}$	$\frac{\text{fa}}{\text{ol}}$	$\frac{\text{fs}}{\text{px}}$
	Sv class		
1662 Hoffmann	**63.0**%	17.8%	15.7%
179 Klytaemnestra	85.0%	17.8%	15.7%
1858 Lobachevskij	32.0%	18.7%	16.4%
2042 Sitarski	76.5%	21.0%	18.0%
2504 Gaviola	**61.3**%	19.5%	16.9%
	Sr class		
17 Thetis	76.6%	18.8%	16.5%
	Sa class		
25143 Itokawa	*21.5*%	30.6%	25.3%
43 Ariadne	*21.1*%	30.7%	25.4%

Table 6: Mineral contents as calculated with Dunn *et al.* 2010 formulas for S-complex objects in the IRTF database. Cells in which the $\frac{\text{opx}}{\text{opx+ol}}$ ratio falls in the range covered by Dunn *et al.* 2010 samples, corresponding to most credible results (see discussion in 8), are indicated in bold. Results for 17 Thetis and 2042 Sitarski could also be considered quite believable, considering that fall quite close to the Dunn *et al.* 2010 sample $(45\%, 70\%)$ range. Results corresponding to the paradoxical range of BAR< 0 are marked in italic as they are the least credible.

Looking at the Sv class, we can see two asteroids with orthopyroxene content within the credible range of the Dunn *et al.* 2010 sample: 1662 Hoffmann with $\frac{\text{opx}}{\text{opx+ol}} = 63\%$ and 2504 Gaviola with $\frac{\text{opx}}{\text{opx+ol}} = 61.3\%$. Result for 2042 Sitarski, $\frac{\text{opx}}{\text{opx+ol}} = 76.5\%$, is sufficiently close to tha Dunn *et al.* 2010 sample range to be considered believable as well. Even higher result of 85% for 179 Klytaemnestra definitely suggests that $\frac{\text{opx}}{\text{opx+ol}}$ is the highest among the objects in Table 6 for this asteroid, but since it falls far outside of the Dunn *et al.* 2010 sample range, the exact value may not be accurate. Analogous conclusion can be made for 1858 Lobachevskij, which is clearly the most olivine-dominant asteroid among the Sr objects from Table 6, but the exact composition derived may be quite off.

The single Sr class object 17 Thetis (see also Figure 22 right), presents a similar case as 2042 Sitarski, so a value $\frac{\text{opx}}{\text{opx+ol}} = 76.5\%$, allowing an error bar of a few percent, seems believable.

Fayalite and ferrosilite content fractions are all quite similar among Sv and Sr objects analyzed: $\frac{\text{fa}}{\text{ol}} \in [17.8\%, 21\%]$, $\frac{\text{fs}}{\text{px}} \in [15.7\%, 18\%]$, implying that all of these objects lie in the same region of the magnesium-iron line.

Two Sa class asteroid 25143 Itokawa and 43 Ariadne stand out from the other objects, as we have already seen on Figure 23. They are both similar to each other, with very low orthopyroxene contents inferred from applying the formula in a suspicious range of BAR< 0. Ironically, all values calculated for 25143 Itokawa are comparable with mineralogical content measured directly from the sample returned to Earth by the Hayabusa mission (see Nakamura *et al.* 2011). The fayalite and ferrosilite contents are probably legitimate, since they are inferred from band positions (see 8), which are quite definite and similar to each other for the two Sa objects and are not affected by temperature and phase angle corrections (see Figure 23). As for the match in orthopyroxene content, we are presumably dealing here with an instance of two errors cancelling each other out, see discussion in section 8.

A closer look into the reflectance spectra data files for the two Sa objects reveals that they were both done at high phase angles:

- 27.1635° for 43 Ariadne, already exceeding the range for which the phase correction formula was fit in Vishnu Reddy *et al.* 2012 (see 7.2)
- A staggering 66.3135° for 25143 Itokawa. It appears that the day of observation $28th$ of March 2001 was around a time of close approach for this NEA, for in the beginning of March the phase angle was very low. Presumably, the incentive was to take a spectrum from close distance. This may also be a factor distorting the correction,

since a case of asteroid spectrum taken at relative proximity implies a huge difference in the situation geometry and light beam paths involved, and the formulas fit for spectra taken from far away may not be applicable.

Clearly, for surveys which would have mineralogical content determination for a large number of S-complex bodies, observing at high phase angles should be discouraged.

9.2.4 Mineralogical analysis of V class objects

As mentioned in section 8, Burbine, Buchanan, and R. P. Binzel 2007 formulas are applicable for V class asteroids. They do not provide any information regarding olivine content, but allow to determine percentage contents of ferrosilite as well as wollastonite among the pyroxenes, which can be used to place these asteroids on the trenary diagram of pyroxenes. This functionality has been implemented into the Morawski 2023-2024 code. Among the two pairs of Burbine, Buchanan, and R. P. Binzel 2007 equations mentioned in section 8, formulas which rely on the position of the first band are used, because the first band is narrower and steeper near the minimum, which makes the determination more precise. Results are shown on Figure 26. The fact that all points lie on a single line should not be surprising, it is a direct consequence of the fact that equations for both ferrosilite and wollastonite content are linear functions of the position of the first band, so they would also linearly depend on each other.

Among the analyzed set, 10 asteroids were found to have very low enstatite content and wollastonite content over 20%, placing them in the cateogry of augites, and the remaining 29 found themselves in the realm of pigeonites (wollastonite content between 5% and 20%. Most of the asteroids have $\frac{\text{wo}}{\text{px}} > 13\%$, with two exceptions: 27343 Deannashea with $\frac{\text{wo}}{\text{px}} = 10.9\%$, $\frac{\text{fs}}{\text{px}} = 45.8\%$ and nameless 26886 with $\frac{\text{wo}}{\text{px}} = 8.5\%$, $\frac{\text{fs}}{\text{px}} = 39.7\%$.

Most V class asteroids in the IRTF datebase share similar orbital and physical properties, as has been discussed in section 5.4. Nevertheless, since a subtle hint for a linear relationship between semi-major axis and diameter had been detected, it was worthwhile to ask if any correlations with the mineralogical contents inferred from the spectra could be made. The semi major axes of the objects are indicated by the color of points on Figure 27. No trend as such can be detected, merely minor clusterings of points with similar semi-major axes and mineral contents. Without dedicating another study to that subject it is hard to say if there is a reason behind it in the dynamical history of the Solar System, or if it is coincidental. Curiously, semi-major axes of both outliers with lowest wollastonite content are very similar to each other.

Coming back to the $a - d$ parameter space analysis from 5.4, we may recall that there was one outlier in terms of asteroid sizes: 4 Vesta with a diameter of 525.4km, where all others had diameters under 10.5km. According to the mineralogical content calculations done in this project, it appears to also reside in the low-calcium region, with $\frac{\text{wo}}{\text{px}} = 15.7\%$, $\frac{\text{fs}}{\text{px}} = 58.3\%$.

Results presented here should be treated with a degree of caution. The same analysis done using band II based formulas from Burbine, Buchanan, and R. P. Binzel 2007 would shift the entire dataset downwards towards the enstatite region, yielding several asteroids classified in the enstatite category $\left(\frac{\text{wo}}{\text{px}} < 5\%\right)$ and no asteroid classified in the augite category. Choosing band I based formulas seemed favorable and more accurate as this band is less wide and the minimum is clearly defined. Also the authors of Burbine, Buchanan, and R. P. Binzel 2007 were cautious about determining the center of the second band, due to low spectral slopes in that part of the spectrum. So some level of error may already be dormant in the band II dependent formulas. Nevertheless, it is not guaranteed that relying on band I center only is optimal, the truth (and hence the mineralogical contents) may lie somewhere in between.

9.3 Olivine absorption bands among K class objects from the SMASS II database

Although the IRTF database was best suited for absorption band analysis, it seemed worth trying to look for other sets of spectra on which the algorithm introduced 6.2 could be tested. Looking back to section 5, we can see that taxonomic class K had many representatives in the SMASS II database and is expected to have a subtle olivine feature. When analyzing with an average smoothing (6.1) window of size $S = 11$, 33 of the K class asteroids had a positive detection of this feature, and most of the detections look convincing. Examples are presented on Figure 28. Detected features lie at the very end of the wavelength range covered in the database, and there is no infrared data to confirm the taxonomic classification nor to exclude contribution from pyroxenes. Shapes of the detected features vary significantly, which may be partially due to the fact that average smoothing distorts the result, partially due to influence of other minerals, or just due to differences in a way in which the three absorption bands corresponding to Fe^{2+} at M1 and M2

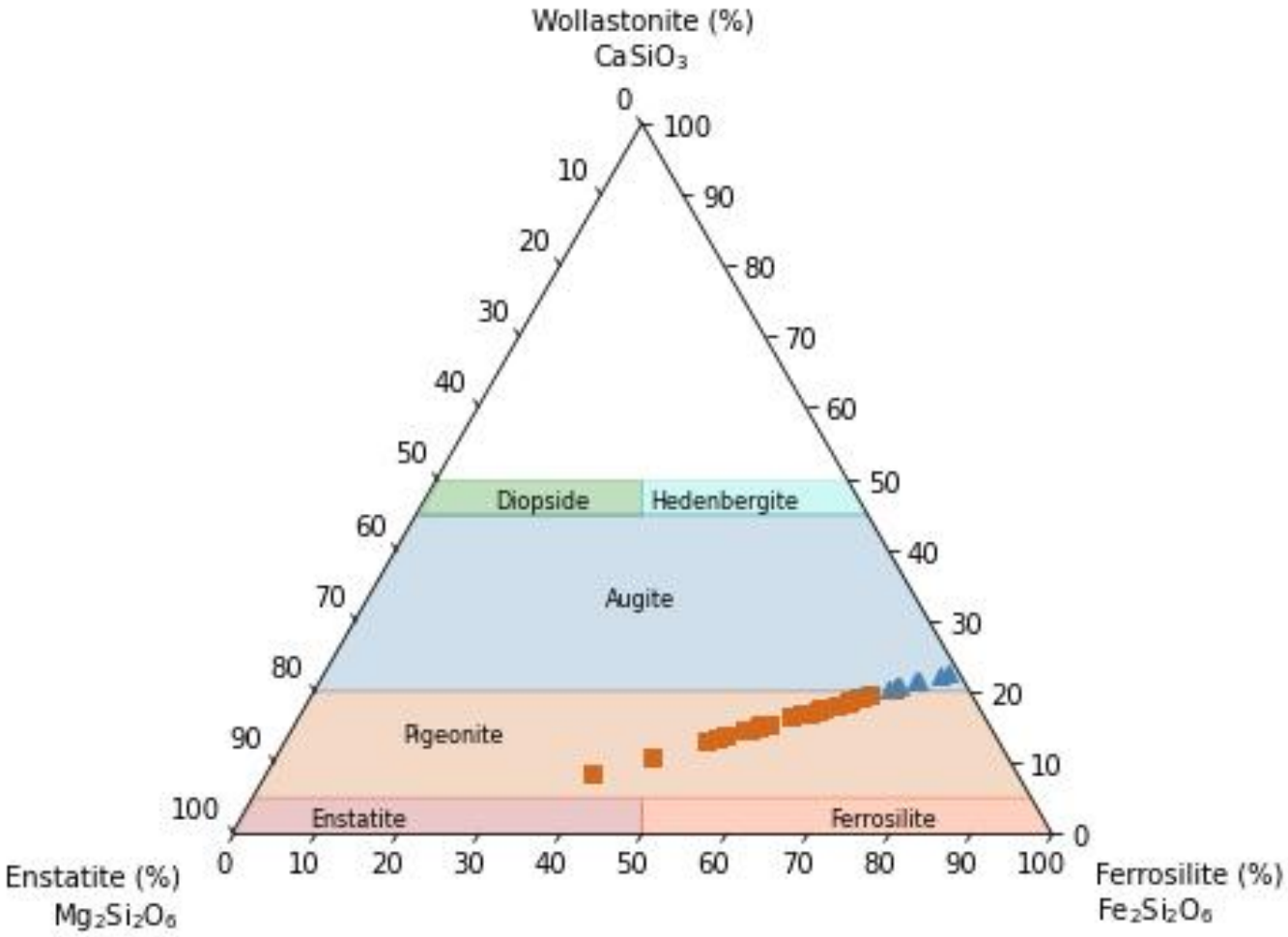

Figure 26: A set of V class asteroids which were included in the IRTF survey database, on the ternary diagram of pyroxenes, as computed with Burbine, Buchanan, and R. P. Binzel 2007 formulas based on positions of the first absorption band in the spectrum, as found with the algorithm introduced in 6.2. Linear alignment of the points is a direct consequence of the linear nature of Burbine, Buchanan, and R. P. Binzel 2007 equations, in other words, only locations along such a line are allowed by these equations. Orange squares correspond to compositions which would be classified as pigeonites, blue triangles would classify as augites (wollastonite content over 20%. The line of possible locations is occupied in a more or less uniform manner in the area of $\frac{\text{wo}}{\text{px}} > 13\%$, and there are two outliers outside of that area, still in the pigeonite category but leaning towards enstatite: 27343 Deannashea with $\frac{\text{wo}}{\text{px}} = 10.9\%$, $\frac{\text{fs}}{\text{px}} = 45.8\%$ and nameless 26886 with $\frac{\text{wo}}{\text{px}} = 8.5\%$, $\frac{\text{fs}}{\text{px}} = 39.7\%$.

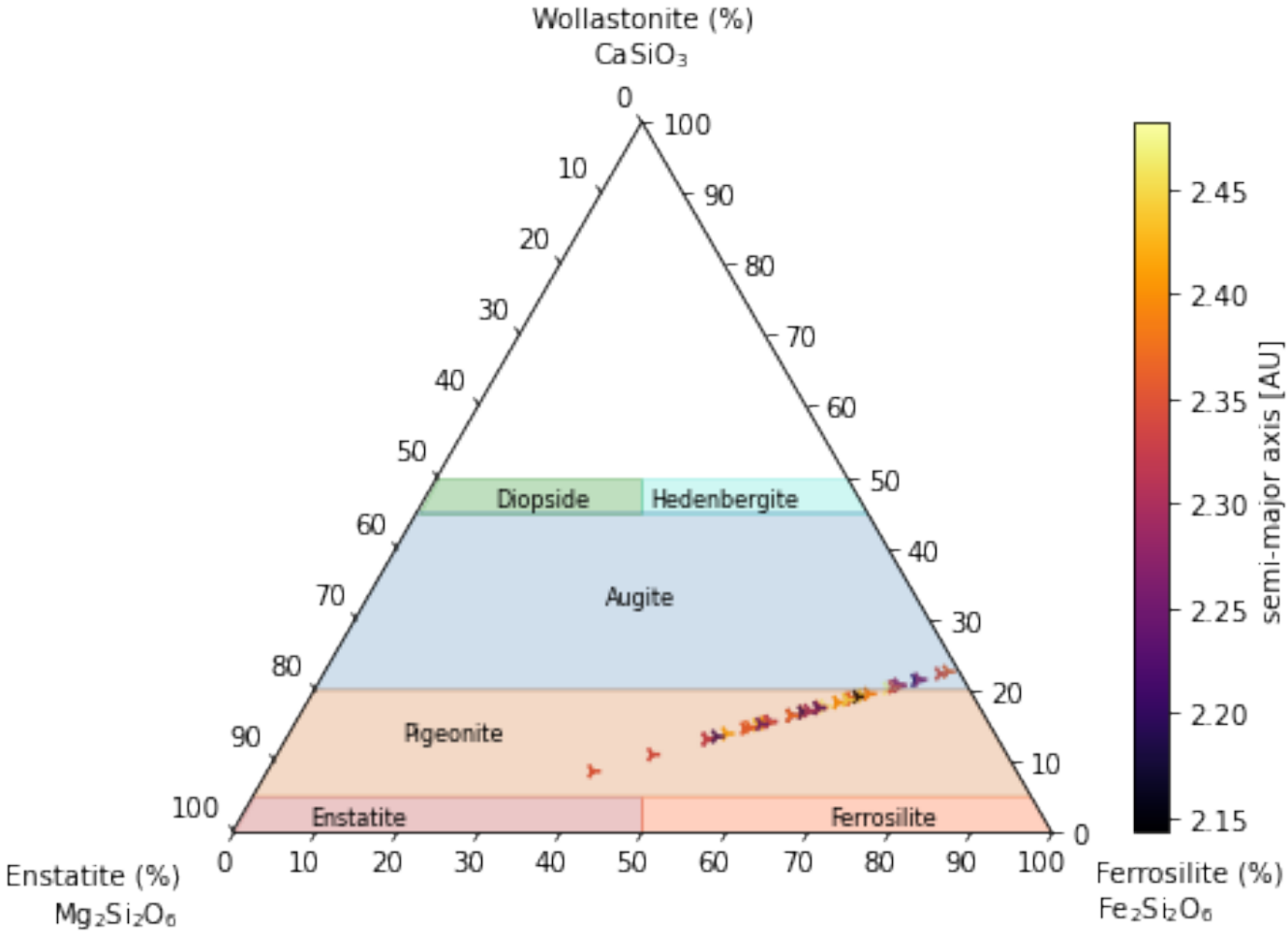

Figure 27: Same as Figure 26, but with color indicating the semi-major axis of the asteroid's orbit. Both low wollastonite outliers seem to be in the same subrange of that parameter, indeed $a = 2.331$AU for 27343 Deannashea and $a = 2.342$AU for object 26886. Outermost objects seem to be close to the pigeonite/augite boundary, on the pigeonite side, but objects which had been categorized as augites are relatively closer to the Sun. Overall, no solid statement about any firm trends could be made.

sites in the crystal lattice come together to form the feature. It is impossible to make any more conclusions within the realm of this project, extensions discussed in 11 could support a more meaningful analysis of the K class spectra.

10 Searching for overlaps - different files with spectra of the same asteroid

One potential data analysis tool, which has been included in the Morawski 2023-2024 code quite early but has not been developed further, could be looking for overlaps between databases and comparing spectra of the same asteroid found in different databases. Ideally, data from different databases should be unified, but there would be numerous challenges related to that. Meanwhile, let's quickly go over some overlaps detected in the beginning of the project.

Within the 5 databases 2.1-2.4 (2.5 has been added much later and not compared with the other databases as there were hardly any overlaps anyways, and other aspects such as much bigger wavelength range and availability of observation dates which set that database introduced in 2.5 apart from others) there are 132 asteroids, which have their spectrum present in more than one (most of these cases are asteroids that were covered both by SMASS and SMASS II, a total of 97 overlap cases between those two). It is interesting to plot these spectra together. It is immediately clear, that while in some cases the spectrum looks almost the same in all databases, in other cases the differences are big.

A frequent occurrence is that two spectra agree up to some wavelength (usually $\sim 0.7\mu$m), then one becomes noticeably dimmer than the other for longer wavelengths, see Figure 29. This seems to almost certainly be caused by phase reddening. However, in order to verify this hypothesis, observation dates would have to be found to calculate phase angles. As explained in 7.4, this is difficult in general and while it could be done for the Fornasier database (2.1) relatively easily, the cases from Figure 29 come from SMASS and SMASS II (2.4), where no reliable method of extracting observation dates seems to be available. Additionally, even if the observation dates could be found, they would only serve to a confirmation of the phase reddening as difference cause in a yes/no manner, for no mathematical models for the phase reddening exist, only an empirical formula for the BAR correction (Vishnu Reddy *et al.* 2012). Therefore, it seems not to be worthwhile to develop the project towards comparing those, until models for the phase reddening had been developed - which would ideally require a dedicated survey where repeated spectroscopic observations of the same set of asteroids would be made at different phase angles, so that functions could be fit that would describe wavelength-dependent drop in the spectrum.

There are, however, other types of mismatches. In some cases, as exemplified by Figure 30, spectra of the same asteroid seem to look completely different in different surveys. Is it due to instrumentation or data reduction techniques? Or perhaps there were actual physical changes between the two observations? While an interesting question, it has not been explored further in this project, since the fundamental issues of taxonomic classification and absorption bamd detection had been prioritized.

Perhaps the most curious is yet another case of discrepancy, see Figure 31. In that case, the spectra align very well for longest and shortest frequencies, and yet there's a range in the middle where one is dimmer. Could it be due to an actual absorption band present only during one observation (e. g. observation of a different side of an asteroid, different conditions on an asteroid depending on it's position at orbit)? Perhaps, but the evidence for the cases shown on Figure 31 did not seem strong enough. In particular the algorithm for detection of absorption bands from section 6 failed to detect an absorption band even in the most suggestive case of 737 Arequipa, despite testing it with different sets of input parameters.

11 Conclusions and perspectives

Considering all the work done and partial results obtained, a conclusion can be made that a lot can be achieved in terms of the defined objective of automating analysis of reflectance spectra of asteroids, but patience is required to differentiate between a variety of special cases and inconsistencies between the databases. An algorithm for detecting absorption bands introduced in 6.2 is definitely a success. However, any conclusions drawn in this report are impaired with a variety of uncertainties, such as:

- Quality of the spectra accessed from different databases differs and some of the results of statistical nature may be tarnished because of that. A way to evaluate the quality of spectra in a programmatic way would have to be included, and parameters which would reject spectra below a certain threshold of some quality measure would need to be introduced in order to try accounting for that

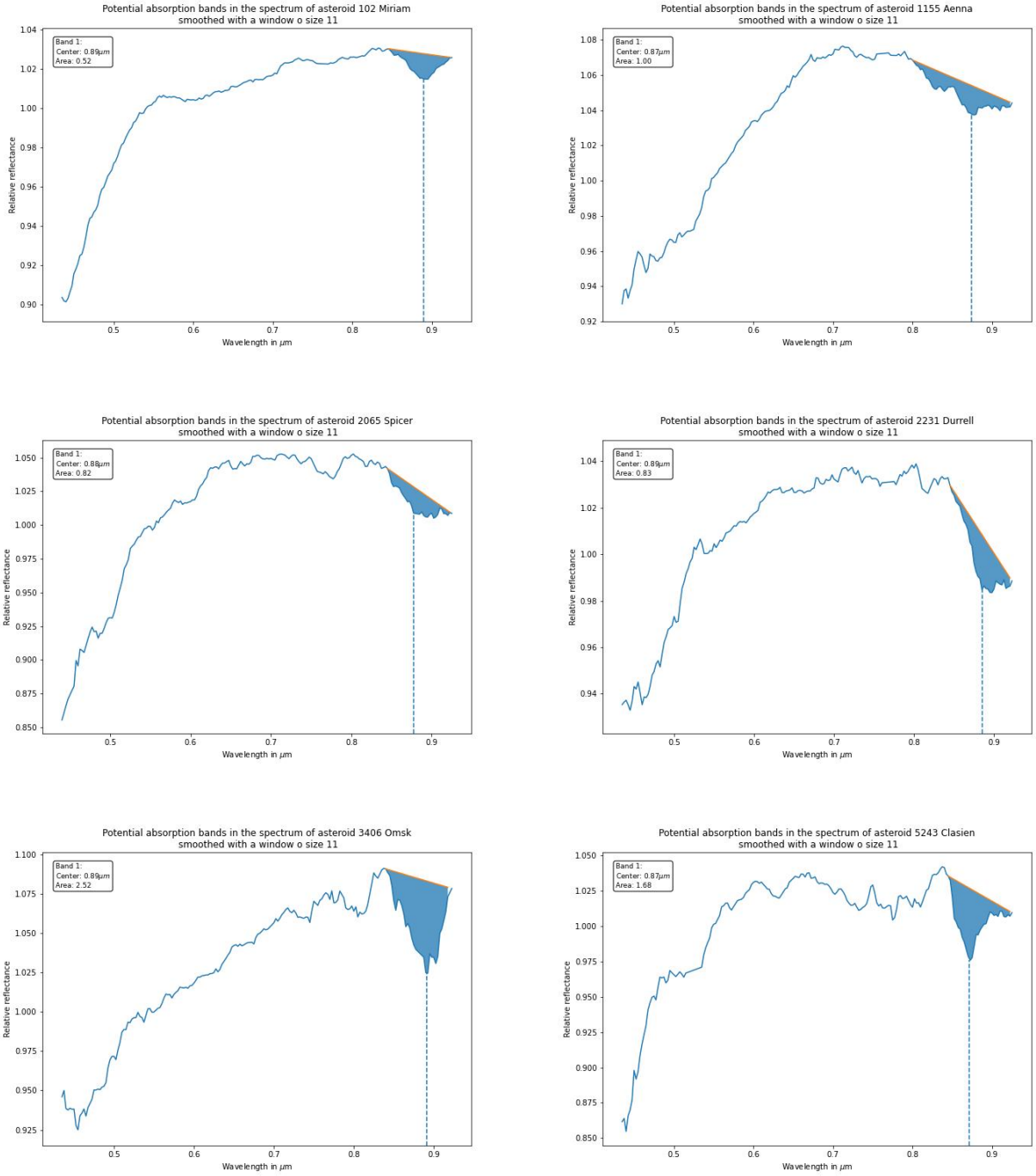

Figure 28: Examples of olivine features detected among the K class asteroids of the SMASS II database (2.4). As we can see, the sizes and positions of the features, as well as continuum slopes differ significantly, but the detections mostly look convincing. Olivine feature is known to be a superposition of three absorption bands corresponding to Fe^{2+} at different sites in the crystal lattice (M1 and M2), whichj is a source of asymmetries in that feature. Further analysis cannot be done until models such as the Modified Gaussian Model (Sunshine, Pieters, and Pratt 1990) will be incorporated in the code.

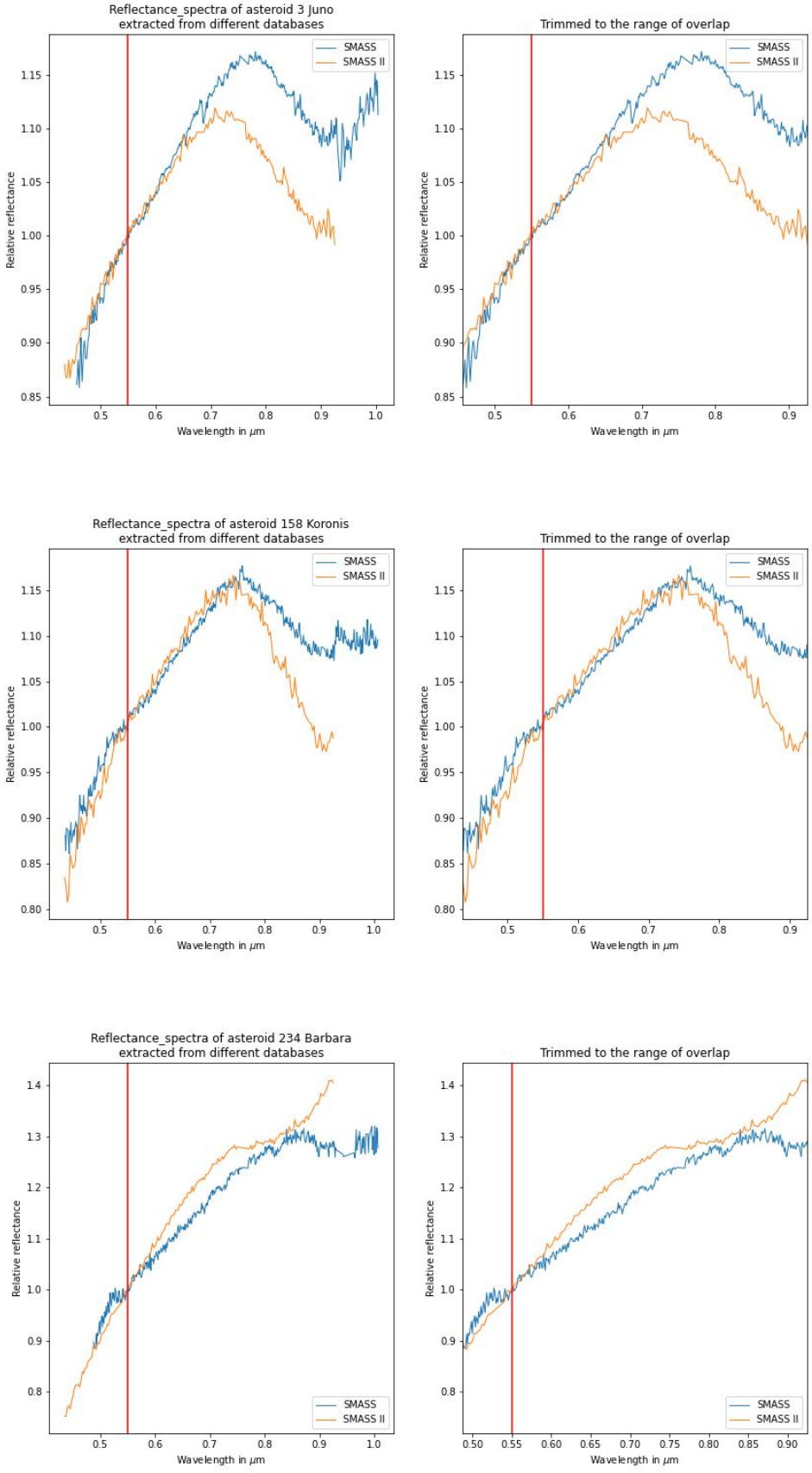

Figure 29: Examples of pairs of spectra with an accordance at shorter wavelengths and a discrepancy at longer wavelengths. A suspected cause would be phase reddening, but needs to be verified.

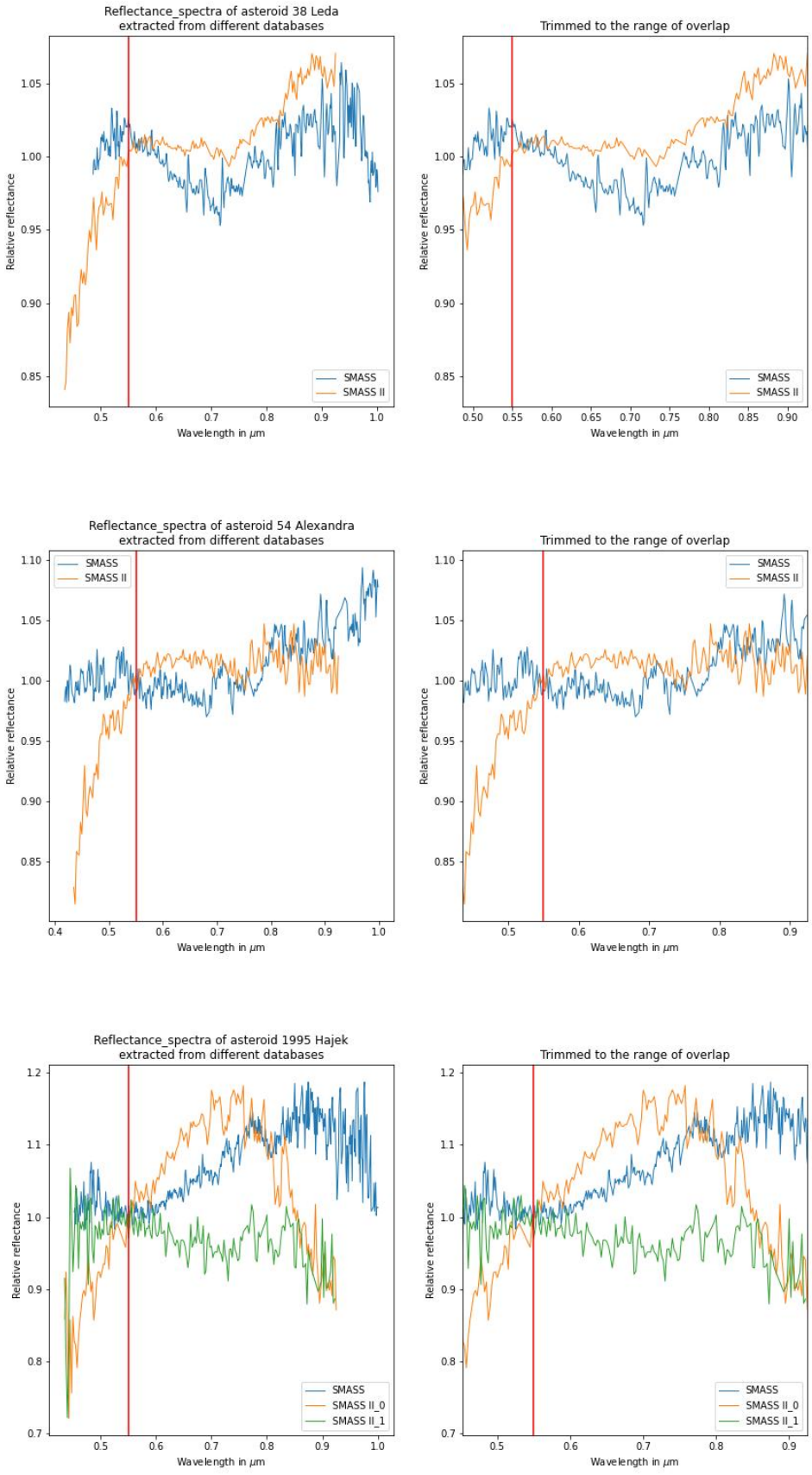

Figure 30: Examples of pairs of spectra with mismatches in the entire wavelength range. It is hard to guess immediately if these differences are of instrumentational, or also physical nature. Curiously, the last case of 1995 Hajek shows not only a big difference between data from two surveys, but also a stark discrepancy between two spectra taken within the same survey (SMASS II_0 and SMASS II_1 in the legend refer to two instances of spectrum from SMASS II database).

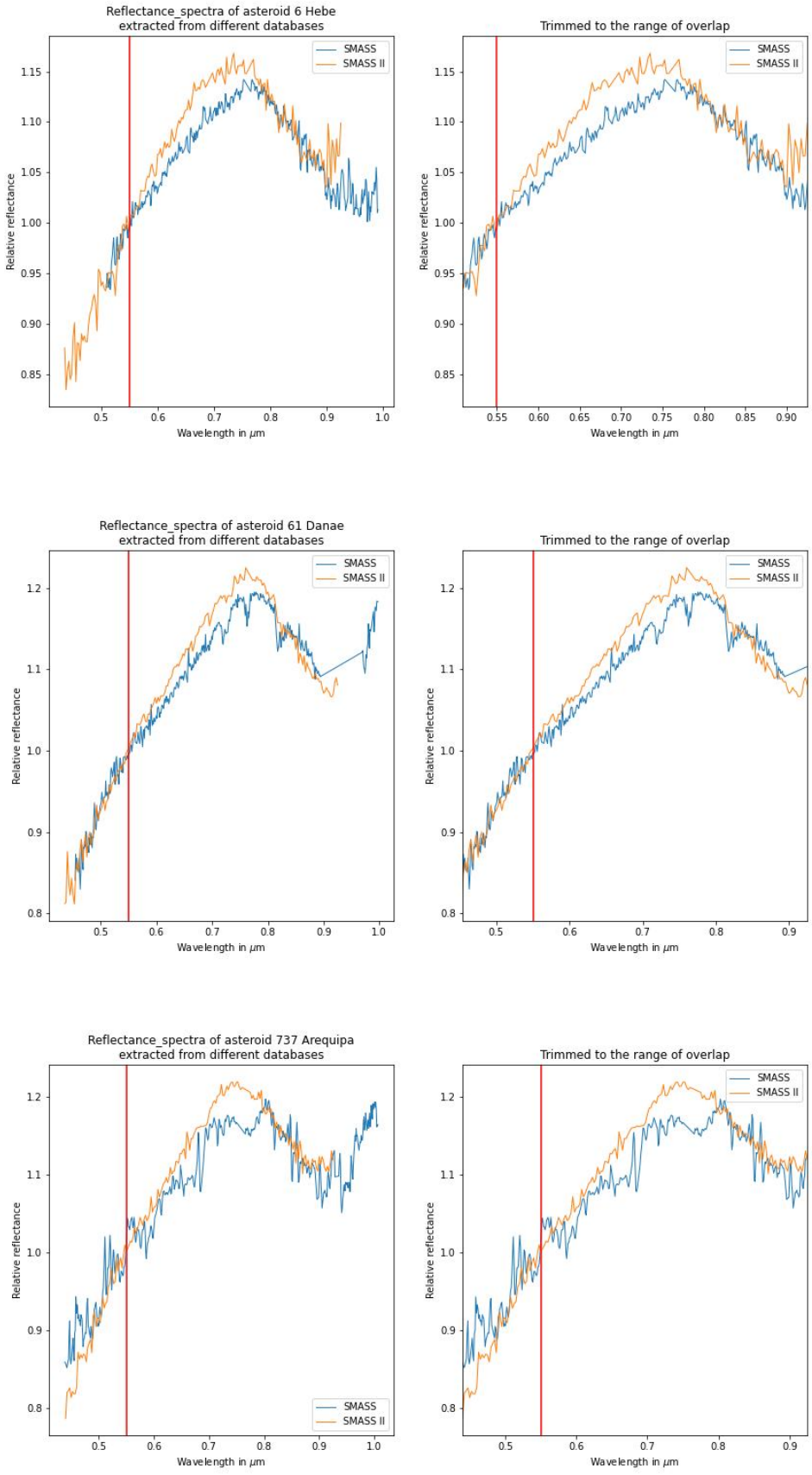

Figure 31: Examples of pairs of spectra with a discrepancy at mid-range wavelengths. Especially the lower plot (737 Arequipa) seems to hint at an absorption feature that would only be present in the first observation.

- Formulas for temperature and phase corrections (7.2) and for mineralogical contents (section 8) have limited ranges of applicability, which were sometimes somewhat stretched in this project. Any derived mineralogical contents should not be given full trust until a number of other surveys come to comparable conclusions, or the mineralogical contents could be examined differently (ideally by sample retrieval by space missions). Further work needs to be done in the realm of deriving empirical formulas from meteorite and asteroid observations, before results of such automated bulk analyzes (which will always inevitably give less attention to individual cases and caveats for each of them) could find mineralogical contents of asteroids in a credible manner.

Overall, the project should be seen more as a search for ways in which large datasets of reflectance spectra of asteroids could be studied in the future, with a focus on accessing data, algorithms, data analysis functions and ways of presenting data on different graphs, rather than work towards robust claims on compositions of asteroids. Quantitative results on those were obtained for some special cases (9.1.2,9.2.3,9.2.4) as a side result, but they were not in the focus, and a number of question marks surrounds the numbers obtained, as explained in relevant sections.

Putting mineralogical conclusions aside, we can point out that there are many more tools which could be employed in the code and could enhance the ability to analyze data and draw meaningful conclusions. Although side quests such as exploring the overlaps mentioned in section 10 could be interesting, the two major tasks which should be undertaken, should the project develop further, are:

1. **Inversion of the Modified Gaussian Model.** The model, introduced in Sunshine, Pieters, and Pratt 1990, aims to describe absorption features as a superposition of modified gaussians, which are meant to correspond to separate bands linked with different energy level transitions. It should be possible to try fitting sums of such modified gaussians to bands detected with the 6.2 algorithm, or to fit the model separately and compare with the results of the algorithm. Some additional supporting information for mineralogical studies could be inferred from deconvolving absorption features into gaussian-like bands, so the idea would definitely be worthwhile.

2. **Fitting the Shkuratov Model.** As introduced in Shkuratov *et al.* 1999, the model represents one of the two approaches to modelling reflectance spectra with radiative transfer theory. As discussed in Morawski 2023, the simplicity and invertibility of the model make it a tool of more potential than the other approach from Hapke 1981, which is why it has been implemented in the CANA library (De Pra *et al.* 2018). It would definitely be interesting to incorporate it in the project. However, it is not that straightforward, as the model depends on optical constants which are only well measured for a handful of minerals (see discussion in V. Reddy *et al.* 2015). This would also make it difficult, if not impossible, to try inferring mineralogical contents from fitting the Shkuratov model alone, since there are no formulas linking the two and the approach in the CANA library is rather that of forward modelling: optical constants for each mineral must be provided in a file and a compositional mixture needs to be defined in the code to generate a modeled spectrum.

Unfortunately, the temporal limitations of this project, together with numerous challenges faced along the way and described in this report, have not allowed, as of now, to include those functionalities.

References

Adams, J. B. (Nov. 1974). "Visible and near-infrared diffuse reflectance spectra of pyroxenes as applied to remote sensing of solid objects in the solar system". In: 79, pp. 4829–4836. DOI: 10.1029/JB079i032p04829.

Bottke Jr., William F. *et al.*, eds. (2002). *Asteroids III*. University of Arizona Press, p. 785.

Burbine, T. H., P. C. Buchanan, and R. P. Binzel (Mar. 2007). "Deriving Formulas from HED Spectra for Determining the Pyroxene Mineralogy of Vesta and Vestoids". In: *38th Annual Lunar and Planetary Science Conference*. Lunar and Planetary Science Conference, p. 2117.

Bus, S. J. (June 2009). "IRTF Near-IR Spectroscopy of Asteroids V1.0". In: *NASA Planetary Data System*, EAR-A-I0046-4-IRTFSPEC-V1.0, EAR-A-I0046-4-IRTFSPEC–V1.0.

Bus, S. J. *et al.* (Sept. 2008). "Bus-DeMeo Taxonomy: Extending Asteroid Taxonomy Into The Near-infrared". In: *AAS/Division for Planetary Sciences Meeting Abstracts #40*. Vol. 40. AAS/Division for Planetary Sciences Meeting Abstracts, 28.22, p. 28.22.

Bus, Schelte J. and Richard P. Binzel (July 2002). "Phase II of the Small Main-Belt Asteroid Spectroscopic Survey. A Feature-Based Taxonomy". In: 158.1, pp. 146–177. DOI: 10.1006/icar.2002.6856.

Chapman, C. R., D. Morrison, and B. Zellner (May 1975). "Surface Properties of Asteroids: A Synthesis of Polarimetry, Radiometry, and Spectrophotometry". In: 25.1, pp. 104–130. DOI: 10.1016/0019-1035(75)90191-8.

Cloutis, E. A. *et al.* (Oct. 1986). "Calibrations of phase abundance, composition, and particle size distribution for olivine-orthopyroxene mixtures from reflectance spectra". In: 91, pp. 11, 641–11, 653. DOI: 10.1029/JB091iB11p11641.

De Pra, M. N. *et al.* (Oct. 2018). "CANA: A Python package for the analysis of hydration in asteroid spectroscopic and spectrophotometric data". In: *AAS/Division for Planetary Sciences Meeting Abstracts #50*. Vol. 50. AAS/Division for Planetary Sciences Meeting Abstracts, 315.02, p. 315.02.

DeMeo, F. E. *et al.* (2015a). "The Compositional Structure of the Asteroid Belt". In: *Asteroids IV*, pp. 13–41. DOI: 10.2458/azu_uapress_9780816532131-ch002.

DeMeo, F. E. *et al.* (2015b). "The Compositional Structure of the Asteroid Belt". In: *Asteroids IV*. Ed. by P. Michel *et al.* Tucson: University of Arizona, pp. 13–41. DOI: 10.2458/azu_uapress_9780816532131-ch002.

Dunn, Tasha L. *et al.* (Aug. 2010). "A coordinated spectral, mineralogical, and compositional study of ordinary chondrites". In: 208.2, pp. 789–797. DOI: 10.1016/j.icarus.2010.02.016.

Dyar, M. Darby *et al.* (Dec. 2023). "A machine learning classification of meteorite spectra applied to understanding asteroids". In: 406, 115718, p. 115718. DOI: 10.1016/j.icarus.2023.115718.

Fornasier, S. *et al.* (Dec. 2010). "Spectroscopic survey of M-type asteroids". In: 210.2, pp. 655–673. DOI: 10.1016/j.icarus.2010.07.001. arXiv: 1007.2582 [astro-ph.EP].

Gaffey, Michael J. *et al.* (Dec. 1993). "Mineralogical Variations within the S-Type Asteroid Class". In: 106.2, pp. 573–602. DOI: 10.1006/icar.1993.1194.

Gartrelle, G. M. *et al.* (Feb. 2021a). "Gartrelle et al. IRTF Asteroid Spectra V1.0". In: *NASA Planetary Data System*, p. 6. DOI: 10.26033/r34k-2238.

Gartrelle, G. M. *et al.* (July 2021b). "Same family, different neighborhoods: Visible near-infrared (0.7-2.45 μm) spectral distinctions of D-type asteroids at different heliocentric distances". In: 363, 114295, p. 114295. DOI: 10.1016/j.icarus.2020.114295.

Hapke, B. (June 1981). "Bidirectional reflectance spectroscopy. 1. Theory". In: 86, pp. 4571–4586.

— (June 2002). "Bidirectional Reflectance Spectroscopy. 5. The Coherent Backscatter Opposition Effect and Anisotropic Scattering". In: 157.2, pp. 523–534. DOI: 10.1006/icar.2002.6853.

JetPropulsionLaboratory (n.d.). https://ssd-api.jpl.nasa.gov/doc/horizons.html#ephem_typeEphemeris look-up with the Horizons API.

JetPropulsionLaboratory, NASA (n.d.[a]). https://ssd.jpl.nasa.gov/tools/sbdb_lookup.html/Small Body Database.

— (n.d.[b]). https://ssd-api.jpl.nasa.gov/doc/sbdb_query.htmlSmall Body Database Query API.

Lauretta, D. S. *et al.* (Oct. 2017). "OSIRIS-REx: Sample Return from Asteroid (101955) Bennu". In: 212.1-2, pp. 925–984. DOI: 10.1007/s11214-017-0405-1. arXiv: 1702.06981 [astro-ph.EP].

Lumme, K., E. Bowell, and A. W. Harris (June 1984). "An Empirical Phase Relation for Atmosphereless Bodies". In: *Bulletin of the American Astronomical Society*. Vol. 16, p. 684.

Mainzer, A. *et al.* (Dec. 2011). "NEOWISE Observations of Near-Earth Objects: Preliminary Results". In: 743.2, 156, p. 156. DOI: 10.1088/0004-637X/743/2/156. arXiv: 1109.6400 [astro-ph.EP].

McMillan, R. S. (Jan. 2000). *Spacewatch Survey of the Solar System*. Technical Report, FRS-305430 Lunar and Planetary Lab.

Michel, P., F. E. DeMeo, and W. F. Bottke, eds. (2015). *Asteroids IV*. University of Arizona Press, p. 895.

Miller, P. (Dec. 2016). "International Astronomical Search Collaboration: Online Educational Outreach Program in Astronomical Discovery for Middle School, High School, & College Students and Citizen Scientists". In: *AGU Fall Meeting Abstracts*. Vol. 2016, ED22A-07, ED22A–07.

Morawski, J. (Dec. 2023). *Mathematical Models of Reflectance Spectra of Asteroids - Theory and Applications*. University of Coimbra, Polo I, Coimbra, 3004-531, Coimbra, Portugal.

— (2023-2024). *Spectra playing*. Jupyter Notebook. Python code developed for the project with markdown documentation.

Moskovitz, Nicholas A. *et al.* (Aug. 2010). "A spectroscopic comparison of HED meteorites and V-type asteroids in the inner Main Belt". In: 208.2, pp. 773–788. DOI: 10.1016/j.icarus.2010.03.002. arXiv: 1003.2580 [astro-ph.EP].

Nakamura, Tomoki *et al.* (Aug. 2011). "Itokawa Dust Particles: A Direct Link Between S-Type Asteroids and Ordinary Chondrites". In: *Science* 333.6046, p. 1113. DOI: 10.1126/science.1207758.

NASA (n.d.[a]). https://science.nasa.gov/mission/psyche/Description of Psyche mission.

— (n.d.[b]). https://science.nasa.gov/dwarf-planets/ceres/exploration/Ceres: Exploration.

PlanetaryDataSystem (n.d.). https://sbn.psi.edu/pds/archive/spectra.htmlAssembly of asteroid spectra.

Popescu, M. *et al.* (July 2019). "Near-Earth asteroids spectroscopic survey at Isaac Newton Telescope". In: 627, A124, A124. DOI: 10.1051/0004-6361/201935006. arXiv: 1905.12997 [astro-ph.EP].

Reddy, V. *et al.* (2015). "Mineralogy and Surface Composition of Asteroids". In: *Asteroids IV*. Ed. by P. Michel and et al. Tucson: Univ. of Arizona, pp. 43–63. DOI: 10.2458/azu_uapress_9780816532131-ch003.

Reddy, Vishnu *et al.* (Jan. 2012). "Photometric, spectral phase and temperature effects on 4 Vesta and HED meteorites: Implications for the Dawn mission". In: 217.1, pp. 153–168. DOI: 10.1016/j.icarus.2011.10.010.

Sanchez, Juan A. *et al.* (July 2012). "Phase reddening on near-Earth asteroids: Implications for mineralogical analysis, space weathering and taxonomic classification". In: 220.1, pp. 36–50. DOI: 10.1016/j.icarus.2012.04.008. arXiv: 1205.0248 [astro-ph.EP].

Shkuratov, Y. *et al.* (Feb. 1999). "A Model of Spectral Albedo of Particulate Surfaces: Implications for Optical Properties of the Moon". In: 137.2, pp. 235–246. DOI: 10.1006/icar.1998.6035.

Singer, R. B. and T. L. Roush (Dec. 1985). "Effects of temperature on remotely sensed mineral absorption features". In: 90, pp. 12, 434–12, 444. DOI: 10.1029/JB090iB14p12434.

Stokes, G. H. *et al.* (1998). "The Lincoln Near-Earth Asteroid Research (LINEAR) Program". In: *Lincoln Laboratory Journal* 11.1.

Sunshine, Jessica M., Carle M. Pieters, and Stephen F. Pratt (May 1990). "Deconvolution of mineral absorption bands: An improved approach". In: 95.B5, pp. 6955–6966. DOI: 10.1029/JB095iB05p06955.

Tholen, D. J. (Sept. 1984). "Asteroid Taxonomy from Cluster Analysis of Photometry." PhD thesis. University of Arizona.

Union, International Astronomical (n.d.). https://minorplanetcenter.net/Minor Planet Center.

van Houten-Groeneveld, I. *et al.* (Oct. 1989). "The 1977 Palomar-Leiden Trojan survey". In: 224.1-2, pp. 299–302.

Vilas, Faith and Michael J. Gaffey (Nov. 1989). "Phyllosilicate Absorption Features in Main-Belt and Outer-Belt Asteroid Reflectance Spectra". In: *Science* 246.4931, pp. 790–792. DOI: 10.1126/science.246.4931.790.

Xu, S. *et al.* (May 1995). "Small Main-belt Asteroid Spectroscopic Survey: initial results." In: 115.1, pp. 1–35. DOI: 10.1006/icar.1995.1075.

Yano, H. *et al.* (June 2006). "Touchdown of the Hayabusa Spacecraft at the Muses Sea on Itokawa". In: *Science* 312.5778, pp. 1350–1353. DOI: 10.1126/science.1126164.