POLITECNICO MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

Executive Summary of the Project

Remote Sensing of celestial rocks: prediction of the chemical structure from spectral analysis

Author: Erica Espinosa, Mattia Gentile, Davide Lo Piccolo, Sophie Retif

Tutor: Matteo Fontana

Course: Nonparametric Statistics

Academic year: 2021-22

Date: January 2022

1. Introduction

As suggested by the paper *"Retrieving magma composition from TIR spectra: implications for terrestrial planets investigations"* by A. Pisello, F. P. Vetere et al. [2], a valuable help in the study of geological evolution of planets is the study of its surface. Currently, however, this kind of information is scarce because the only celestial bodies of which we can have surface samples are the Moon and Mars. To overcome this problem, it is proposed an interpretation of spectra measured on glassy rocks, representing different magmatic series and showing variability in SiO_2 content and in other alkali (Na_2O and K_2O) content. The aim of this project is to identify what are the critical characteristics of spectra and find proper models to estimate the amount of SiO_2 and alkali in a substance given these information coming from emissivity and reflectance. The reason why we focused our attention on the estimation of silica and alcali is because these two are the building blocks of the so called TAS diagram (Figure 1), an important tool used by geologists to classify rocks based on their chemical composition. After having achieved such knowledge, it will be possible to study the geological history of a solid body even in situations where it is materially impossible to reach it. To do so, we first analyzed our curves to extract the main features, then we proceeded to perform analysis on the effect that temperature has on these features and, after that, we built Generalized Additive Models and prediction intervals.

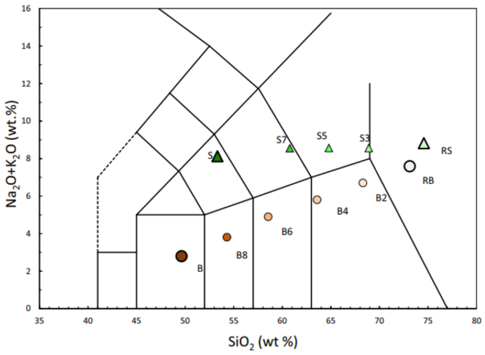
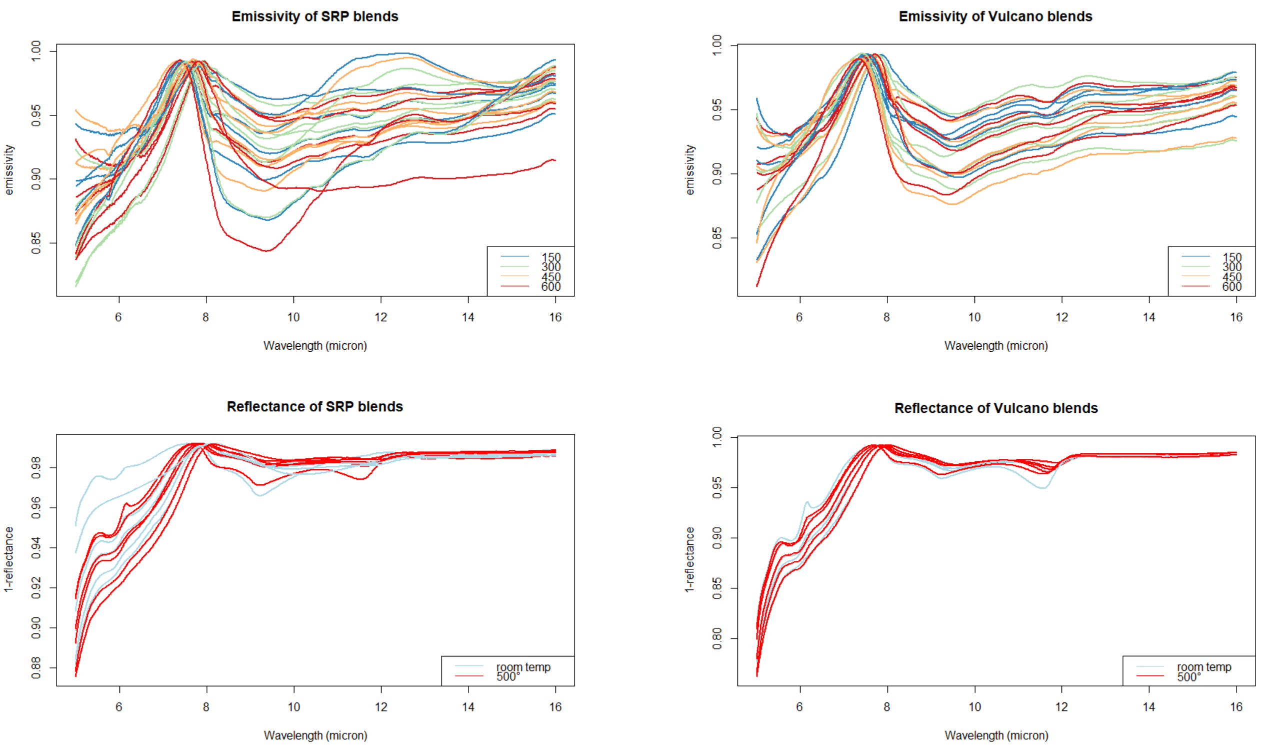


Figure 1: Theoretical TAS diagram and the collocation of our data.

2. Dataset

The data we had to study were the spectrograms of emissivity and reflectance of different magmatic series, showing variability in SiO_2 content and in alkali content, and measured at differ-


Emissivity of SRP blends

Emissivity of Vulcano blends

Reflectance of SRP blends

Reflectance of Vulcano blends

Figure 2: All our dataset represented, divided in, at the top, the spectrograms of the emissivity and at the bottom the spectrograms of the reflectance; at the left the graphs for the SRP blends and at the right the ones for the Vulcano blends.

ent temperatures. The materials were synthesized starting from four natural rocks collected at the Island of Vulcano (Italy) and at Snake River Plain (USA). The first two of them, collected in Italy, are shoshonite and rhyolite while the two rocks collected in the USA are basalt and rhyolite. The four rocks were crushed, pulverized and molten different times in order to obtain homogeneous glassy materials and then mixed together, obtaining, in this way, two series of products with intermediate compositions. For the Vulcano series, shoshonite and rhyolite were mixed in five different weight proportions 100:0, 70:30, 50:50, 30:70 and 0:100 named respectively S, S7, S5, S3 and RS. In the same way, basalt and rhyolite were composed with the following proportions: 100:0, 80:20, 60:40, 40:60, 20:80 and 0:100 named B, B8, B6, B4, B2 and RB. After having synthesized all the blends, spectra for each one of them have been measured in the Thermal Infrared Range (TIR) from 7 to 16 μm at different temperatures. Reflectance measurements were performed on our experiment samples at room temperature of 20°C and at 500°C, while emissivity measurements were performed at 150°C, 300°C, 450°C and 600°C. All the spectra composing our dataset are represented in Figure 2.

3. Features selection

All the spectra show a peak of maximum emissivity (minimum reflectance) at $\sim$8 μm, where emissivity is close to the unity and reflectance is close to zero. This spectral feature is related to Christiansen effect and the wavelength at which it occurs is called Christiansen Feature (CF). Under the suggestion of A. Pisello, F. P. Vetere et al. [2], we identified it, along with the value of emissivity measured at CF (CFval), as an important feature to include in our models as it differs significantly between blends. Another feature we have assumed to be relevant is the Transparency Feature (TF), namely the wavelength at which there is a minimum (or maximum in the case of reflectance) in the interval preceded by CF; again, we considered also the value of emissivity measured in correspondence of TF (TFval). The last feature we wanted to take into account is the first derivative of the spectrogram. In fact, it can be seen from Figure

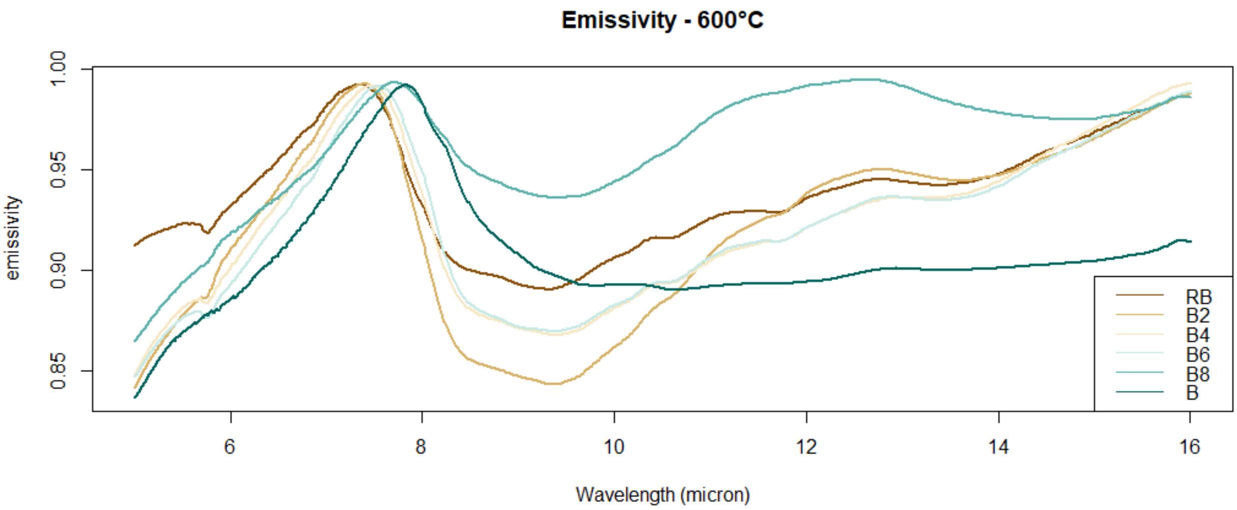


Figure 3: This plots represent the spectrogram of emissivity of the SRP blends at 600°C, it is visible how the slope changes from one blend to another.

3 that the different blends have different slopes in the region of the curves following TF. We thought that it could be helpful to consider the norm L^2 of the derivatives as a feature in our future models. Finally, in order to build a regression and prediction model, we used as response the actual values of the percentages in weight of silica and alkali for each blend.

4. Preliminary analysis

Before building our predictive models for silica and alkali we performed few exploratory analysis on the extracted features and on data themselves. In particular, we moved in two main directions: a priori outlier detection, and inspection of the effect that temperature has on our features. Regarding the former, we employed both functional boxplot and outliergram in order to detect any sort of curve with a particular behaviour. What we came up with is that B600 (i.e. blend B measured at 600°C) could be identified as a amplitude outlier, according to the output of the functional boxplot. As we can see from Figure 4, the main difference that this curve has with respect to all the others is a peculiar flat behaviour for the wavelengths after the transparency feature. However, in the end, we decided not to label this datum as an outlier but keep it into consideration for further analysis; this mainly because of the scarcity of data we were provided with and also because, in this specific framework, it is inappropriate to talk about outliers. Indeed, data were built and measured ad hoc for the purpose, and any sort of particular behaviour means more variability, valuable information in data and not erroneous measurements.

Coming to the analysis on temperature, as we can see from Figure 5, we noticed an interesting behaviour. After the Christiansen Feature, at a lower value of temperature corresponds a higher mean value of emissivity and a less pronounced Transparency Feature. Such information looked meaningful to be studied and so we performed different kinds of permutation tests, both with a multivariate approach and a functional one. We decided to focus our attention only on data coming from emissivity because reflectance did not show any kind of difference between high and low temperature curves. In the multivariate framework we divided our data in groups according to the temperature at which they were measured; then, we considered each feature one at time, and we tested the equality of their distribution between each pair of groups. To do so we took into consideration three different test statistics:

- Difference between sample means,
- Difference between sample medians,
- L^2 norm of the differences between the first and third quartiles.

For example, we took the difference between the sample mean of Christiansen Feature for data

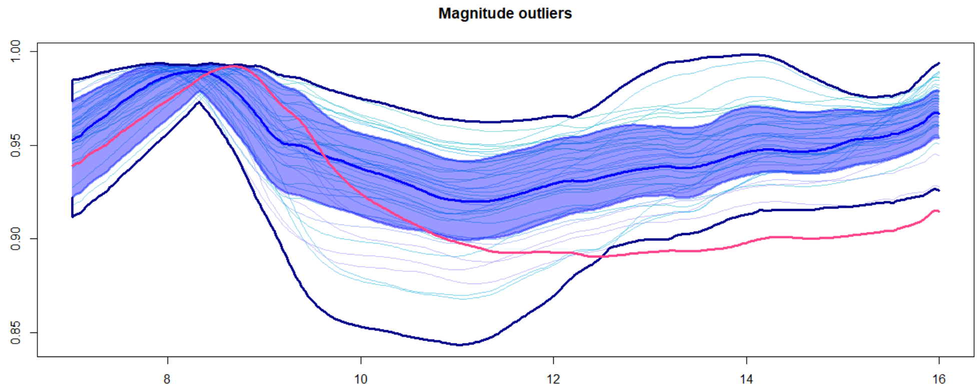


Figure 4: Functional BoxPlot of all our data. It can be seen that B600 is labeled as an outlier because its trend, after the TF, does not vary as much as all the other data.

at 150°C with the sample mean of Christiansen Feature at 600°C as null test-statistic. Then we applied a permutation strategy, sampling from the pooled data and computing, at each iteration, the value of the test-statistic on permuted data. Last, we compared the vector of new T-statistics with our T_0 and obtained the p-value of the test. After all these tests, we obtained statistical evidence to affirm that, when looking at the value of Transparency Feature, data measured at 450°C and 600°C are different from the other two groups, in accord to what we expected just looking at the curves. To achieve a further validation, we also employed a more global approach, i.e. functional permutation test. This allowed us to compare not just a single feature at a time, but the whole behaviour of curves with each other and detect in which specific regions we have a substantial difference. As shown in Figure 6, the interval of wavelengths for which we have that temperature groups differ is the one around the Transparency Feature, confirming, again, what we obtained previously. As a conclusion, we were pretty sure that the effect of temperature on features, in particular on the value of Transparency Feature, had to be taken into consideration. To achieve this task, we introduced a dummy variable distinguishing “low

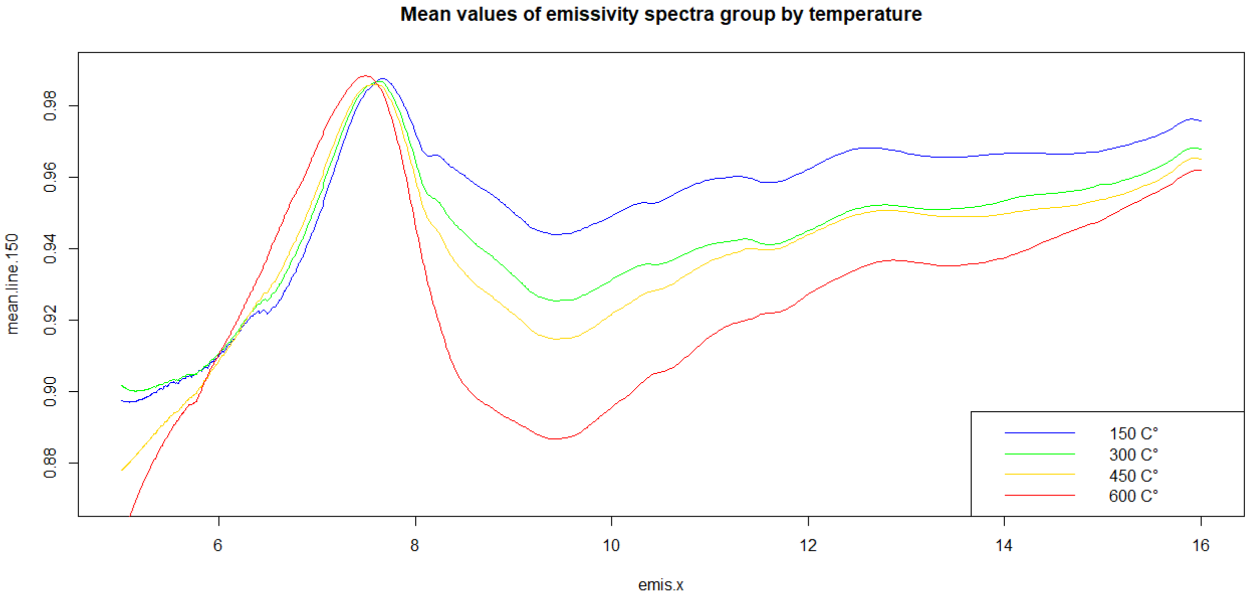


Figure 5: Mean of the spectra of emissivity of all the blends grouped by temperature.

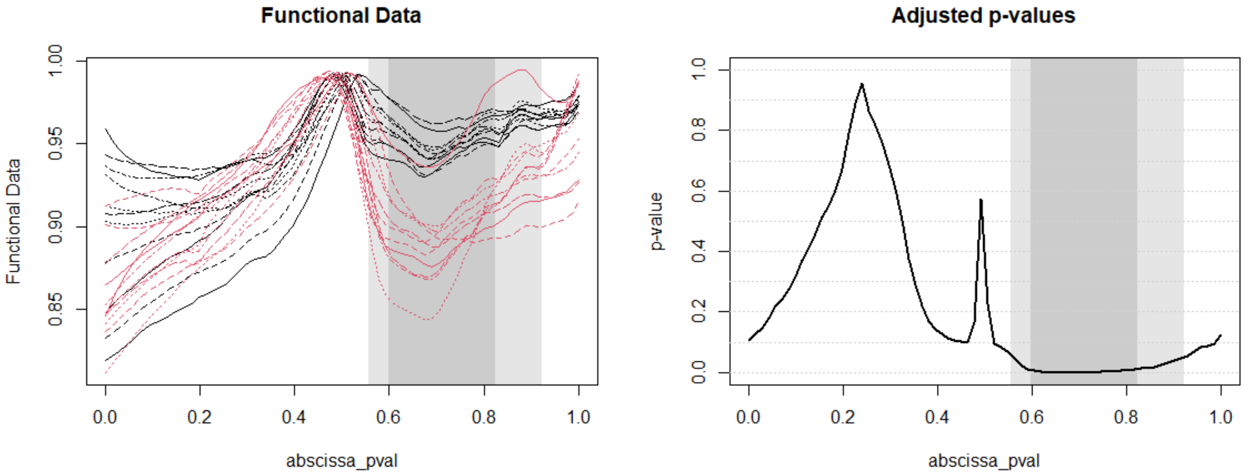


Figure 6: Results of the functional permutation test among the emissivity spectra measured at 150°C and 600°C. The groups of functions show a significative difference in the neighborhood of the Transparency Feature.

temperatures" (150°C and 300°C) from high ones (450°C and 600°C).

5. Regression

After some preliminary analysis and data exploration carried out to have a deep understanding of our dataset, we proceeded to build our predictive models. As described in the section related to features selection, we noticed a strong relationship between the curves of emissivity and reflectance and the percentage of both silica and alkali in our blends. In particular, the main features extracted from the curves that give information about chemical composition are the Transparency Feature and, mainly, the Christiansen Feature. Further insights coming from preliminary analysis, induced us to create also a dummy variable to keep track of the two temperature families and use it in interaction with the emissivity value of TF. On the other hand some variables that have been identified during features selection did not result in significant regressors for our models; this is the case of the derivative that, while slightly improving just one of the models (silica explained through reflectance), was totally useless for the majority of the cases. Not to introduce too many different quantities, we decided to discard this feature.
In order to put together all these meaningful information we decided to employ a Generalized Additive Model that allowed us to control smoothness of the predictor functions and to deal with nonlinear relationships.
The percentage of silica is thus predicted by two different models. The first one takes into account information about emissivity while the second one deals with reflectance data. The models are the following:

$$\begin{aligned}(\mathrm{SiO_2})\% = \beta_{0,e} + \beta_{1,e}\mathrm{CF}_{\mathrm{wave},e} + \\ +f(\mathrm{CF}_{\mathrm{value},e}) + f(\mathrm{TF}_{\mathrm{wave},e}) + \\ +f(\mathrm{TF}_{\mathrm{value},e}) + \epsilon\end{aligned}$$

$$\begin{aligned}(\mathrm{SiO_2})\% = \beta_{0,r} + f(\mathrm{CF}_{\mathrm{wave},r}) + \\ +f(\mathrm{TF}_{\mathrm{value},r}) + \epsilon\end{aligned}$$

Here, the $\mathrm{CF}_{\mathrm{wave},e}$ and $\mathrm{CF}_{\mathrm{value},e}$ denote respectively the value of the wavelength and the value of the emissivity at which the Christiansen effect occurs. While $\mathrm{TF}_{\mathrm{wave},e}$ and $\mathrm{TF}_{\mathrm{value},e}$ are related to the wavelength and the emissivity value of the Transparency feature. The notation is similar for the covariates regarding the reflectance measurements, using r instead of e in the subscript.
It is worth noticing that we were able to simplify our model by applying a linear relationship between some regressors instead of using a smooth term. Indeed, the value of the wavelength of the Christiansen Feature in the model that uses the emissivity data shows a strongly negative linear dependence. To validate such simplification, we

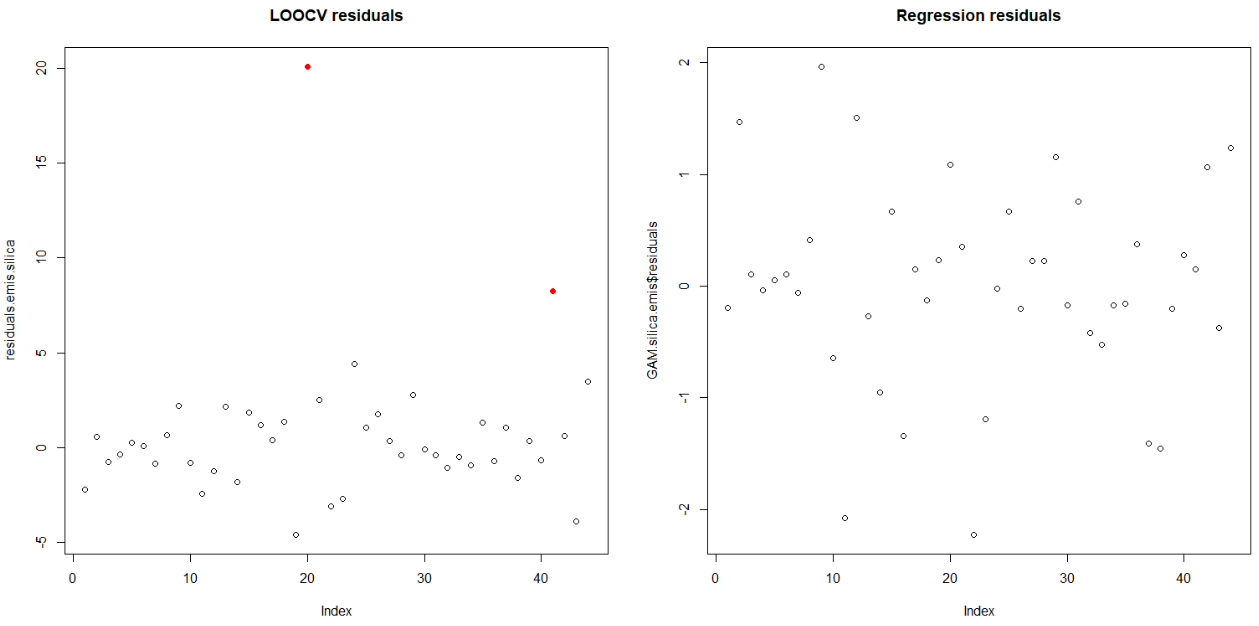


Figure 7: Residuals of the model Silica ∼ f(emis) obtained with the Leave-One-Out cross-validation. It shows two anomalies and, not considering the latter, the range of the residuals is wider (the values range from -5 to 5).

performed an ANOVA test to compare the regression with the linear term and the one with the smooth term. The p-value of the test denoted statistical evidence to state that the two models were equally efficient and that allows us to keep the simplest model between the two.
On the other side, the smooth functions used for the $\mathrm{CF}_{\mathrm{value}}$, $\mathrm{TF}_{\mathrm{wave}}$ and $\mathrm{TF}_{\mathrm{value}}$ exploit cubic spline basis defined by a modest sized set of knots spread evenly through the covariates values.
In an analogous fashion, the two models built for alkali are the following:

$$\begin{aligned}(\mathrm{Na_2K_2O})\% = \beta_{0,e} + \beta_{1,e}\mathrm{CF}_{\mathrm{wave},e} + \\ + f(\mathrm{CF}_{\mathrm{value},e}) + f(\mathrm{TF}_{\mathrm{wave},e}) + \\ + f(\mathrm{TF}_{\mathrm{value},e}) + \epsilon\end{aligned}$$

$$\begin{aligned}(\mathrm{Na_2K_2O})\% = \beta_{0,r} + f(\mathrm{CF}_{\mathrm{wave},r}) + \\ + f(\mathrm{TF}_{\mathrm{value},r}) + \epsilon\end{aligned}$$

Also in this case the smoothing function f is a cubic spline with knots spread evenly through the covariates values, while the linear effect of $\mathrm{CF}_{\mathrm{wave}}$ is, again, strongly negative.
A separate discussion has to be done for the application of the dummy variable in the GAM model. We recall that such variable discriminate "low temperatures" (`dummy` = 0) from "high temperatures" (`dummy` = 1). In a preliminary work it seemed very influential and very effective in the predictive models, in particular in the regression of silica percentage using emissivity. Nonetheless, by including also the dummy variable, the model was reaching a really small bias and that was a clear sign of a possible overfitting behaviour. If we add this fact to the very high adjusted R^2 that we obtained in each model, it looked mandatory to corroborate our models and to do so we performed leave-one-out cross validation. In particular, we compared the residuals obtained from such procedure with the ones of our regression looking for potential overfitting. Indeed, due to the limited amount of data we had, our predictions result to be quite sensitive to overfitting and we had to reduce some of our models (ex: eliminate temperature effect). In particular, the models explaining silica through emissivity resulted to be the most overfitted model (and also the one with the largest number of parameters). In Figure 7 it is visible how the residuals obtained by the model do not coincide with the one obtained through LOOCV, as in the latter they range in a larger interval. At the end we ended up with the models previously described that represent the right compro-

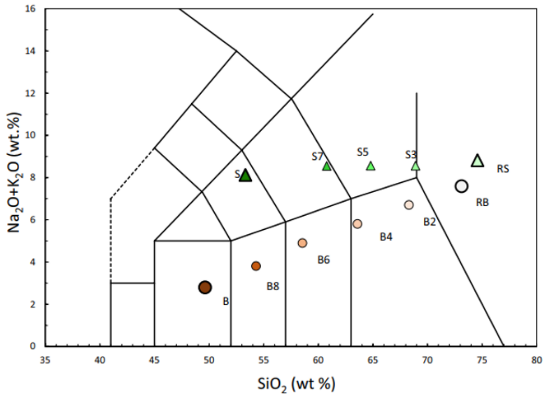


Figure 8: TAS diagram with the collocation of our blends.

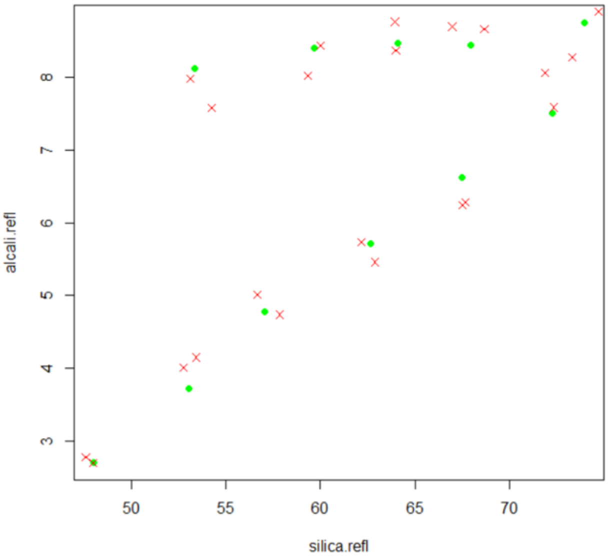


Figure 9: The red crosses represent our predictions of the silica and alkali values of our samples with respect to their reflectance measured at 20°C and 500°C, while the green dots are the real values. We can notice that the predicted values quite correspond to the real ones.

mise between bias and variabiliy of the prediction. The comparison between the real values of the composition of our blends and the values obtained through our models is depicted in Figure 9.

6. Conformal Prediction

As we already said, the main goal of our analysis is to make inference on the chemical composition of the rocks from the emissivity and reflectance spectra, so after doing pointwise forecasting, we built prediction intervals using a conformal approach. In particular, we chose as non-conformity measures the absolute value of regressor residuals and we set α equal to 0.05. Moreover, we decided to make our prediction on the measurements of the dataset, so that, in this way, we have been able to compare the results we obtained with the true values. In Figure 10, the prediction intervals are represented as a line with blue dots on theirs extremes and the true values as red dots. Obviously the larger is the number of red dots contained in their respective interval, the more precise are the prediction intervals. As we can see from the two pictures, most of prediction intervals contains the real value of silica and alkali percentage and this fact attests a good quality of our predictive model.

7. Conclusions

The study of spectrograms to quantify the glassy composition of the materials under study could be a good tool for studying the magmatic origin of celestial bodies that cannot be reached physically. After various analyses of our data, we extracted the features of greatest interest, performed permutation tests to identify whether temperature had a significant effect on the selected features, and found regression and prediction models for the SiO_2 and alkaline components in our materials. Given the scarcity of data, we had to simplify our models so that the number of parameters to be estimated did not exceed the size of our sample. Among the selected features we noticed that the most relevant are the Christiansen Feature and the Transparency Feature and the corresponding spectrogram values. In addition, overfitting problems arose (forcing us to further reduce selected features). Finally, we obtained fairly good models

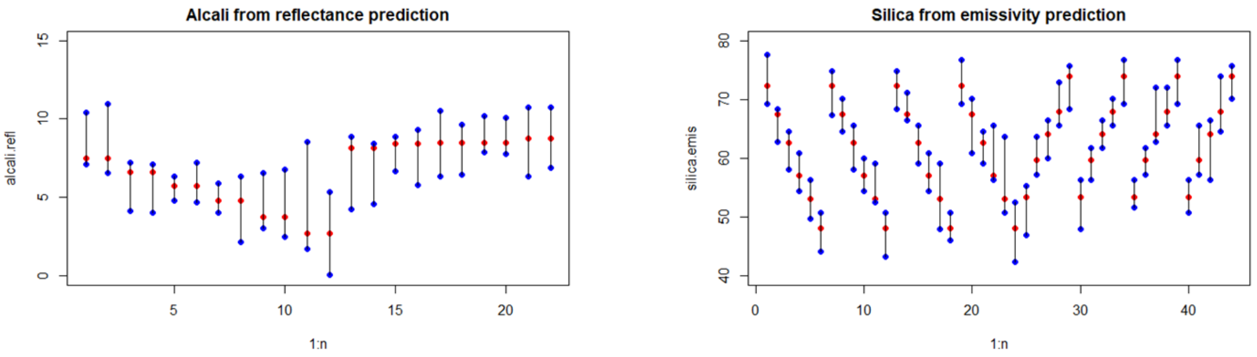


Figure 10: Confidence interval from the confomal prediction with confidence level $1-\alpha$ equal to 0.95.

for the regression and prediction of silicon and alkaline content based on both emissivity and reflectance. We were able to construct appropriate prediction intervals in the TAS diagram and the actual glass content values fall within these intervals. It would be very interesting to be able to continue the study with more data and with blends having a smaller difference in glassy material content between blends, which could help us solve the overfitting problem.Another possible development for this work would be to create a model that uses features extracted from both emissivity and reflectance at the same time, in order to predict the silicon and alkaline content of a given material. Summing up, we did manage to identify characteristics of spectra providing important information about the chemical composition of our blends but for many reasons we could not employ them all for building a model. We can say, that this is for a sure a field of research worth the investment. The path we followed looks promising to us and with a larger amount of data at disposal we will be able to understand better what celestial bodies are made of.

References

[1] B. L. Cooper, J. W. Salisbury, R. M. Killen, and A. E. Potter. Midinfrared spectral features of rocks and their powders. *Journal of Geophysical research*, 04 2002.

[2] Alessandro Pisello, Francesco P. Vetere, Matteo Bisolfati, Alessandro Maturilli, Daniele Morgavi, Cristina Pauselli, Gianluca Iezzi, Michele Lustrino, , and Diego Perugini. Retrieving magma composition from tir spectra: implications for terrestrial planets investigations. *Scientific Reports*, 10 2019.