Mem. S.A.It. Suppl. Vol. 26, 112

Memorie della
Supplementi

Validation of statistical clustering on TES dataset using synthetic Martian spectra

G. Liuzzi[1], F. Mancarella[2], S. Fonti[2], A. Blanco[2], T. L. Roush[3], G. Masiello[1], C. Serio[1], J. R. Murphy[4], and M. Chizek[4]

[1] Scuola di Ingegneria, Università degli Studi della Basilicata, Via dell'Ateneo Lucano, 10, 85100 Potenza (PZ), Italy, e-mail: giuliano.liuzzi@unibas.it
[2] Dipartimento di Matematica e Fisica "Ennio De Giorgi", Università del Salento, Via Arnesano, 73100 Lecce (LE), Italy
[3] NASA Ames Research Center Moffett Field, CA, USA 94035-1000
[4] Astronomy Department, New Mexico State University (NMSU) Las Cruces, NM, USA

Abstract. In this work we present some results concerning the analysis of Thermal Emission Spectrometer (TES) data, looking at the methane Q-branch spectral signature at 1304 cm^{-1}. Such analysis has been enabled by producing some synthetic spectral datasets, simulating the atmospheric and surface variability observed on Mars, excluding the high latitude regions. The use of synthetic spectra is aimed to provide a better comprehension of the influence that the atmospheric state vector and its composition have on the spectral behavior. This effort is important, because the TES data are characterized by a low resolution (10 cm^{-1}) and a significant random and systematic noise which could, in principle, give results whose quality needs to be improved. We apply statistical clustering of the synthetic spectra to evaluate the effectiveness of detecting methane, and estimating its abundance.

Key words. Mars: methane – Mars: atmosphere modeling – Mars: radiative transfer – Thermal Emission Spectrometer – Data analysis: Statistical clustering

1. Introduction

The discovery of Martian atmospheric methane started a very lively discussion in the community, due to the fact that methane, on Earth, is strongly linked to the presence of life forms, even if a purely geological origin is also possible. However such a discovery implies the current production of the gas, that suggests a geologically active planet. In spite of subsequent identification of methane (Mumma et al. 2009; Fonti & Marzo 2010), the results of such observations have been questioned (Zahnle et al. 2011).

In spite of TES' low spectral resolution, we can take advantage of the large number of spectra per Martian Year available (up to 6 millions). This allows us to use statistical techniques, like clustering, to analyze the datasets. The use of synthetic datasets, if generated with a set of well-calibrated parameters, provide an essential mechanism for interpreting the clustering results on real TES datasets.

We begin by describing the general characteristics of the TES spectra; we then describe in detail the parameters used to produce the synthetic spectra. We show the results of the clustering procedure on the synthetic datasets, looking at the methane detection issue. Finally, the future developments are briefly outlined.

2. TES spectra modeling

2.1. TES spectra: brief overview

The Thermal Emission Spectrometer (TES) is a Michelson interferometer on board of the Mars Global Surveyor (MGS) mission. Its spectral coverage goes from ∼ 6 to ∼ 50 μm (201.6 to 1654.3 cm^{-1}) (Christensen et al. 2001). The nominal spectral resolution is 10 cm^{-1} for the large majority of spectra, and 5 cm^{-1} for a smaller portion of data. However, the real resolutions are 12.5 cm^{-1}, and 6.25 cm^{-1}, because of the self-apodization due to the misalignment of each of the six detectors with respect to the optical axis of the instrument (Christensen et al. 2001).

Fig. 1 is an example of TES emissivity spectrum at 10 cm^{-1}. Because of the low resolution, the only spectral feature which is clearly visible in every spectrum is the atmospheric CO_2 absorption band centered at 668 cm^{-1}; other spectral signatures are visible, with different intensities, in any spectrum, such as the water vapor rotational band features and the large suspended dust band centered at ∼ 1080 cm^{-1}, while the spectral continuum is modeled by the surface emissivity, particularly in the regions between 200 and 500 cm^{-1}, and between 1200 and 1600 cm^{-1}. The methane absorption spectral feature is located in an atmospheric window region. This implies that its detection could be disturbed by the surface spectral features, that we have to model as accurately as possible.

2.2. Martian atmosphere model

The synthetic spectral datasets have been produced by using the MODTRAN radiative transfer code, ver. 3.1, which is a line-by-line model, based on the HITRAN 96 database version (Rothman et al. 1998). In fact, the adoption of a moderate resolution code is quite suitable for reproducing low resolution spectra, like those provided by TES. The simulated spectra are produced at a resolution of 1 cm^{-1}, and then convolved to the TES resolution and sampling. The generation of a dataset of synthetic spectra is based on a look-up table of vertical temperature-pressure profiles. Not the entire TES dataset has been used to analyze the methane concentration in the Martian atmosphere (Fonti & Marzo 2010): in the same way, we have taken into consideration only

Table 1. Pressure grid used in the Martian atmosphere model.

Layer #	Av. pressure (mb)	Av. altitude (m)
1	13.5000000	0.00
2	12.9137000	2981.80
3	10.0572000	6033.60
4	7.8325550	9024.76
5	6.1000000	11956.5
6	4.7506850	14829.9
7	3.6998370	17646.2
8	2.8814360	20406.6
9	2.2440650	23112.1
10	1.7476790	25763.8
11	1.3610940	28362.8
12	1.0600210	30910.2
13	0.8255452	33406.9
14	0.6429352	35854.0
15	0.5007185	38252.5
16	0.3899599	40603.3
17	0.3037011	42907.4
18	0.2365227	45165.7
19	0.1842040	47379.1
20	0.1434582	49548.5
21	0.1117254	51674.8
22	0.0870000	53837.6
23	0.0678000	55822.8
24	0.0528000	57937.1
25	0.0411000	59883.2
26	0.0320000	61918.0
27	0.0249000	63632.8
28	0.0194000	65498.6
29	0.0151000	67487.0
30	0.0118000	69209.2

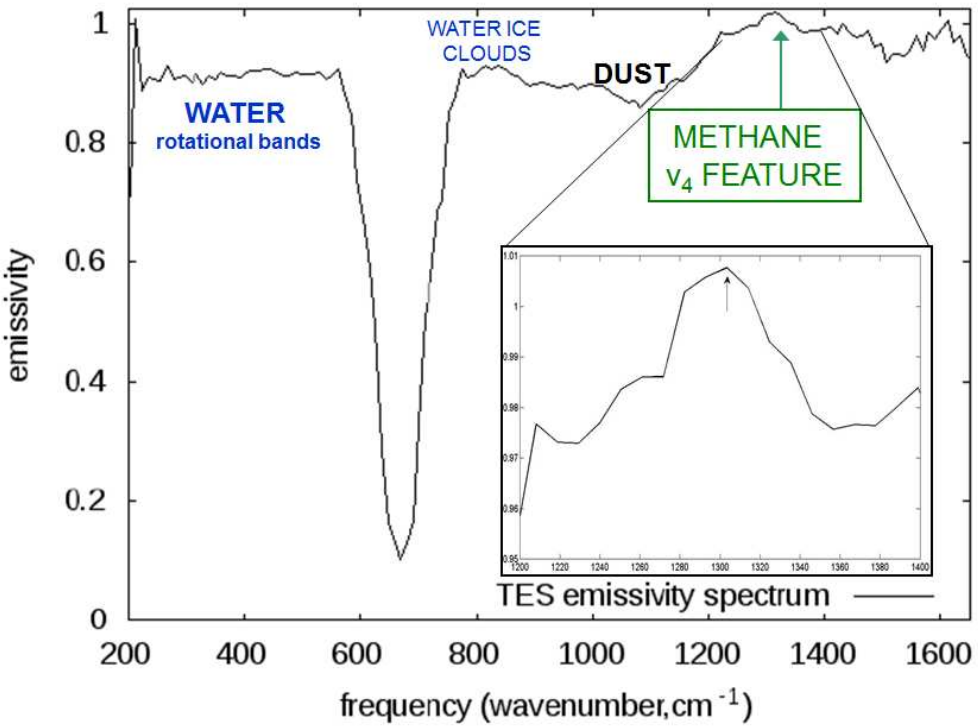

Fig. 1. An example of TES emissivity spectrum. The main spectral features are indicated. In the bottom right inset, a zoom on the region where the CH_4 ν_4 Q-branch is.

the atmospheric state vectors[1] corresponding to particular conditions and locations: from -50 to +50 degrees in latitude, taken in the central part of the day (local time from 11 to 15), and characterized by a nadir-viewing geometry. In addition, for this work, we have selected only profiles corresponding to low-dust concentration in the Martian atmosphere. This choice is justified by the experimented difficulty to retrieve the methane total columnar amount from dusty TES spectra. This, as the other choices generating synthetic datasets, are aimed to represent the entire range of atmospheric and surface variability observed on Mars.

Table 1 shows the pressure grid used as input to produce the synthetic spectra. For convenience, it is the same as is used in the TES temperature profile retrieval algorithm, with the layer boundaries chosen such that the gas within the layer can be considered homogeneous. The grid is extended up to 0.01 mb, which corresponds to ~70 km of altitude, with the zero-level fixed at 13.5 mb [2]. Trials made producing synthetic spectra corresponding to average Martian atmospheric conditions, compared to real TES spectra, have shown that the errors introduced by limiting the number of layers to 30, and by extending the layering up to 0.01 mb, are negligible along all the TES spectral coverage, at the TES low resolution.

The temperature profiles in the look-up table have been retrieved by the TES team using the Conrath algorithm (Conrath et al. 1998), based on the inversion of the atmospheric radi-

[1] "Atmospheric state vector" is intended as the temperature-pressure vertical profile.

[2] Assuming an atmospheric vertical scale of 11.6 km

ance

$$R_\nu^{atm} = \int_{z_s}^{z_t} B[\nu, T(z)] \cdot \frac{\partial\tau(\mu,\nu,z)}{\partial z} dz \qquad (1)$$

on the spectral channels corresponding to the strong CO_2 absorption band at 668 cm^{-1}. The temperature profiles have been considered as representative of the whole atmospheric variability, and used without modifications.

At this first stage, the atmospheric composition has been assumed to be constant with altitude: in other words, the various gases concentrations are the same for each atmospheric layer. This hypothesis makes sense with the CO_2; but is not realistic for the water vapor, or methane. However for the methane the mechanisms for its origin and subsequent destruction in the atmosphere are not known. So there is little information to constrain its abundance with altitude. As a result we leave it well-mixed in all layers.

A different approach is adopted to model the aerosol dust and the water ice vertical profiles. For dust, Heavens et al. (2011) describe the vertical profile of dust as a combination of a Conrath profile (Conrath et al. 2000), which decreases with altitude, and a Gaussian, at a typical altitude of 25-35 km. This combined profile is more general than the simple Conrath profile, so we have decided to use it, yielding:

$$q = q_{Conrath} + q_{Gauss} = q_0 \cdot \exp\left[n(1-\alpha^{-1})\right] + \\ + \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(z-z_l)^2}{2\sigma^2}\right] \qquad (2)$$

where q_0 is the dust mass mixing ratio at z=0, σ the standard deviation (in altitude) of the dust enriched layer, z_l is the centroid (in altitude) of the gaussian distribution, n is a diffusion parameter, which controls how steeply the dust concentration decreases as z increases, and:

$$\alpha = \frac{p - p_{top}}{p_s - p_{top}}, \; or \quad \alpha = \exp(-z/H) \qquad (3)$$

where p_{top} is the atmospheric pressure at the top of the grid (Table 1), p_s the atmospheric pressure at ground, and H is the vertical scale of the atmosphere. Typical values for the parameters above are $\sigma = 7km$, $z_l \approx 30km$, while $n = 0.01$ during the dust storms, and 10-40 times higher during the rest of the Martian Year. Furthermore, the enriched layer is really significant only in the dust storms periods; for this reasons, in the model the Conrath profile component is dominant, while the gaussian enriched layer amplitude has been set 10-15 times lower than the Conrath one, setting $n = 0.25$, and varying q_0 from 10^{-6} to $2.5 \cdot 10^{-5}$. Some examples of dust vertical profiles (with altitude) are given in Fig. 2.

Several observations (by rovers and spacecrafts) have confirmed that ice clouds commonly occur in the middle atmosphere, even in the equatorial regions, where the ice condenses on dust grains. Moreover, many spectra exhibits an absorption feature centered at ~800 cm^{-1}, which can be directly attributed to water ice particles; as a consequence, this atmospheric component cannot be neglected. Water ice absorption is here modeled multiplying the atmospheric emissivity spectrum, obtained by including gases and dust as previously described, by the water ice particles emissivity, scaled with the ice optical depth. Even if this component should not influence the spectrum in the methane band region (around 1300 cm^{-1}), we include it for completeness.

2.3. Martian surface emissivity model

The last spectral component to be modeled is the surface emissivity. To make simulations faster, and to avoid the introduction of redundant hypotheses about the surface composition and emissivity, we have multiplied each simulated atmospheric spectrum by a linear combination of the three average emissivity spectra, representative of the typical Martian surface types: andesitic, basaltic, and dusty, having a different visual albedo (Bandfield & Smith 2003). The three average emissivity spectra are plotted in Fig. 3, in the TES spectral range.

3. Synthetic spectra clustering

3.1. Introduction

Our motivations in creating a synthetic spectral dataset are to: 1) evaluate if the statistical

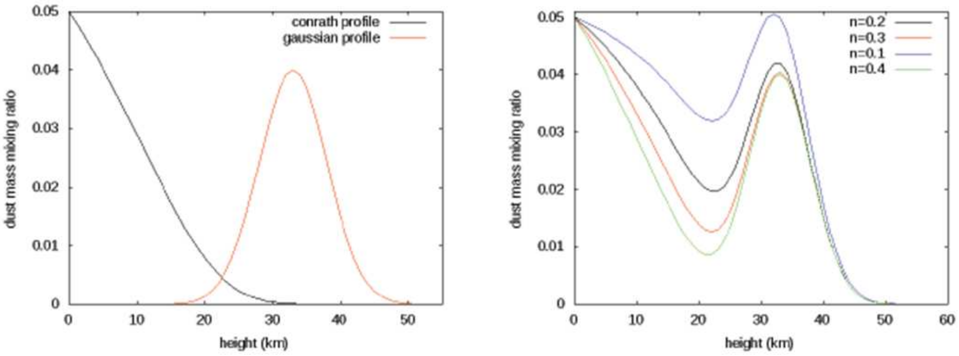

Fig. 2. Typical behavior of the dust mass mixing ratio with altitude. Left panel: the Conrath component and the gaussian component, separated; right panel: sum of the two components, for different values of *n*.

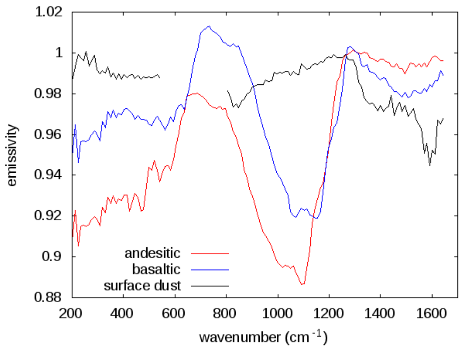

Fig. 3. The three average surfaces spectra used in the model: andesitic, basaltic, and dusty (Bandfield & Smith 2003).

clustering technique used in Fonti and Marzo (2010) can identify methane from a controlled sample; and 2) quantify the minimum methane mixing ratio detectable from a large number of spectra. The minimum number of TES spectra used in the analysis of methane detection and abundance on Mars (Fonti and Marzo 2010) is of the order of 10^4. Because of the computationally intense nature of creating the synthetic spectra, approximately 4 hours for 10^4 spectra, we created a synthetic data set with a number of spectra close to this minimum, approximately 10^4.

3.2. Synthetic spectra generation

The model described in section 2 has been used to produce sets of synthetic spectra. To generate a single spectrum, the temperature vertical profile and the surface pressure have been randomly extracted from the look-up table, while the H_2O vertical profile and the CH_4 mixing ratio have been made varying according to the knowledge that we have about the martian atmosphere composition (see Table 2). Finally, the dust vertical profile and total extinction has been modeled according to the trend introduced in section 2.2. In addition, not all the spectra are characterized by a methane mixing ratio different from 0. In particular, we have chosen to create datasets with approximately 60% of spectra without methane, while the residual 40% of spectra is characterized by

Table 2. Variability of the main parameters used to produce synthetic spectra datasets (*: Atmospheric dust extinction efficiency at 670 nm, measured in km^{-1}).

Parameter	**Min. val.**	**Max. val.**
Surface skin temp.	230 K	300 K
Surface pressure	0.5 mb	11 mb
H_2O Mixing ratio	0 ppmv	800 ppmv
CH_4 Mixing ratio	0 ppbv	80 ppbv
Dust extinction effic.*	0	0.05
Ice tot. optical depth	0	0.2

a methane columnar amount within the limits in Table 2. Both the CH_4 and H_2O columnar amounts have been extracted randomly from a triangular distribution peaked on the average of the max. and min. limits in Table 2.

At this first stage, no random noise has been introduced in the spectra, because in the CH_4 band region that we are looking at, the random noise is much less significant than the variations of emissivity introduced by the surface variability and atmospheric suspended dust.

3.3. Clustering results

The cluster analysis should group the spectra according to their methane content; In an ideal case, the statistical analysis of the input spectra would produce two clusters of spectra; one containing methane (labeled as "methane cluster") and one without methane ("no-methane cluster"). In order to make the clustering process efficient, and to focus the statistical analysis on methane, we have defined a suitable parameter to give in input to the clustering algorithm, the CH_4 band depth, calculated as follows:

$$BD = \frac{\epsilon_{CONT_{SX}} + \epsilon_{CONT_{DX}}}{2 \cdot \epsilon_{CENTER}} \tag{4}$$

where $\epsilon_{CONT_{SX}}$ ed $\epsilon_{CONT_{DX}}$ are the emissivity values in the two channels alongside the methane spectral channel (~1294 and ~1315 cm^{-1}), while ϵ_{CENTER} is the emissivity in the methane spectral channel (~1304 cm^{-1}). In fact, this quantity could give information about the energy absorbed by the methane in the atmosphere, and so it is suitable to be used as input for the clustering algorithm.

Fig. 4 shows the scatter plot of the clustering results; this chart is built putting on the x-axis the emissivity value at 1294 cm^{-1} (the channel on the left of the methane channel), and on the y-axis the emissivity value at 1304 cm^{-1} (methane channel).

A better visualization of the result is provided by having a look to the centroids of the two clusters, in Fig. 5. It can be seen that no obvious methane bands are visible, neither in the "no-methane" cluster, nor in the "methane" one. However, using the approach of Fonti and Marzo (2010) by dividing one cluster by the other, then a feature appears near the methane band as shown in the bottom of Fig. 5. This permits us to label the clusters methane and no methane.

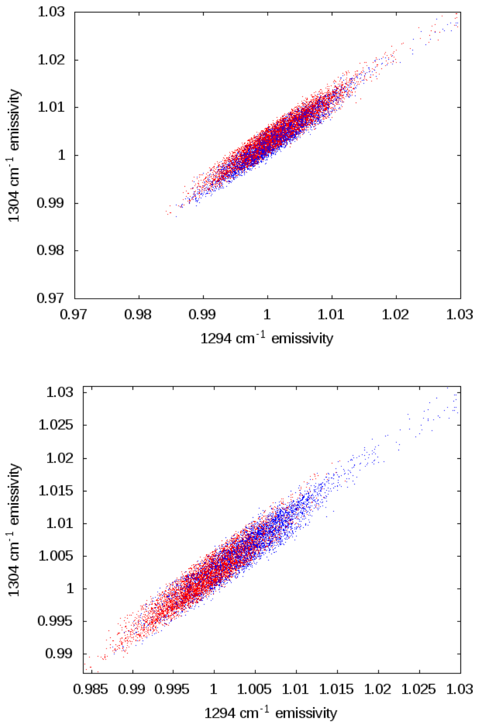

Fig. 4. Top panel: synthetic dataset analyzed; the red points represents the spectra with no methane, while the blue points represents the spectra with a methane mixing ratio different from 0. Bottom panel: result of the clustering. The red points represents the spectra inserted in the "no-methane" group, while the blue points represents the spectra put in the "methane" group.

However, it is apparent from Fig. 4 that there is significant overlap in emissivity values between these two clusters. This is quantified in Table 3 where the number of spectra in each group is defined pre- and post-clustering. Spectra with and without methane are associated with the inaccurate cluster. To quantify this, known the methane abundance of every spectrum, it is easy to calculate the aver-

Table 3. Clustering results.

	Real data	**Clustering results**
# sp. per cluster	5892 no-methane, 3928 methane	4648 no-methane, 5172 methane
ppbv CH_4 per cluster	No-methane: 0 ppbv Methane:(39.5 ± 26.4) ppbv	No-methane: (8.76 ± 18.11) ppbv Methane: (22.14 ± 27.14) ppbv

age methane abundance in each of the groups pre- and post-clustering. The results are in the last row of Table 3. Pre-clustering the group without methane yields an appropriate zero methane abundance value, while after clustering the methane abundance has increased. For the groups containing methane, the estimated abundance is decreased by clustering, likely due to the inclusion of spectra not containing methane in this group.

3.4. Spectra simulation issues

Despite of the criticity introduced by the surface emissivity variability, and particularly by the surface dust, the clustering succeeds partially to distinguish the spectra according to their methane mixing ratio. However, it must be pointed out that the simulation must be improved, especially in the surface component.

Looking at Fig. 6, it is apparent the distributions are different. While there may be a number of possible explanations for this difference, we suggest two explanations here (Mind that the real TES dataset taken in example is made by almost 200000 spectra).

Firstly, our treatment of the variability of the surface emissivity may not be sufficient, especially for the surface dust component. We calculated the average of all the synthetic spectra and divided this by the average of all the TES spectra represented in Fig. 7 and the result is compared to the average Martian dusty surface spectrum (Bandfield & Smith 2003). The overall slope is similar, even though some of the maxima and minima are shifted. This suggests to us that the influence of the surface dust needs to be more accurately accounted for.

Secondly, no random, or correlated, noise was included in calculations of the synthetic spectra while these are inherently present in the TES data.

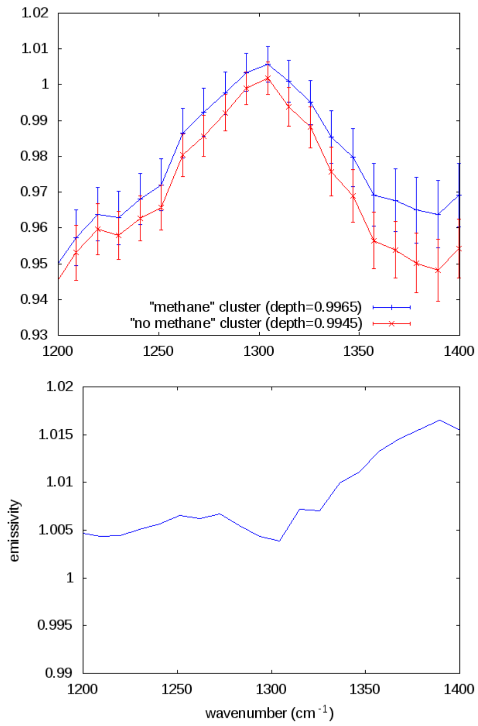

Fig. 5. Top panel: centroids of the two clusters in the spectral region around the methane band; the centroid of the "no-methane" cluster is in red, the centroid of the "methane" cluster is in blue; the error bars are not referred to the instrumental noise, but they represent simply the standard deviation of the spectral emissivity of each cluster. Bottom panel: ratio between the two centroids; the methane band is visible.

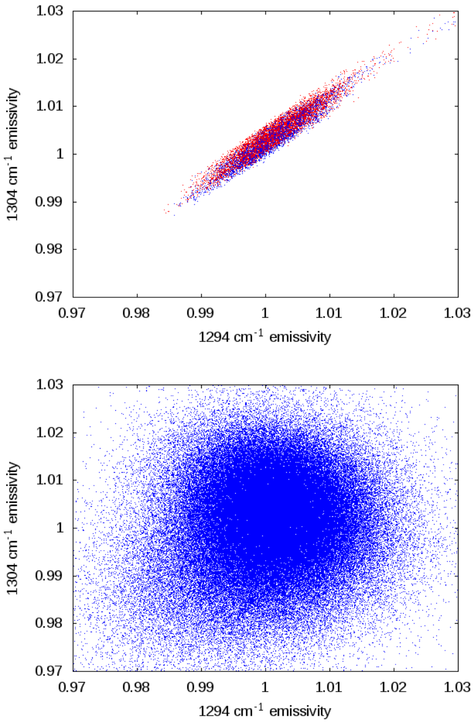

Fig. 6. Top panel: synthetic dataset scatter plot (like in Fig. 4. Bottom panel: example of real TES dataset scatter plot. It is clearly visible that they have a different symmetry.

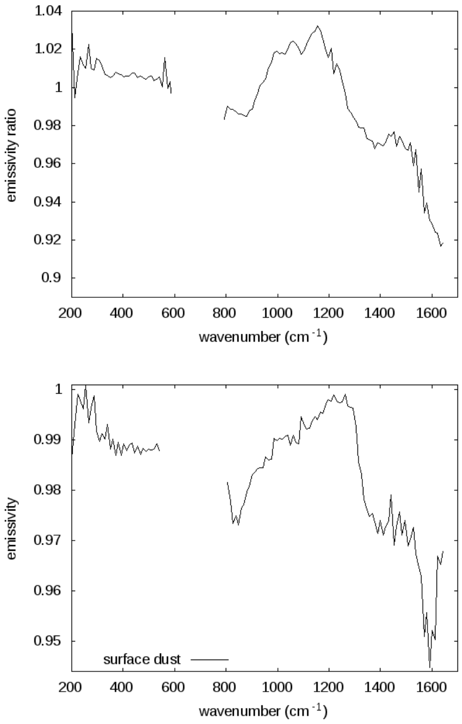

Fig. 7. Top panel: ratio between the average spectra of the two datasets in Fig. 6. Bottom panel: dusty surface average emissivity spectrum.

4. Conclusions and future developments

The results we have shown in this paper describe how it is possible to rely on the clustering technique in analyze TES data for methane retrieving and estimating, even if the spectral resolution is quite poor. In a certain way, this work has been done from a new perspective, which consists in testing the data analysis procedure on synthetic data, whose characteristics are well-known. Regarding the methane detection, we have shown that a straightforward approach consists in using the band depth as clustering parameter, and that it works well enough for our purposes. Of course, the simulation of the surface emissivity must be improved, as well as the atmospheric profiling.

Another aspect that must be improved in the future concerns the radiative transfer model by which we produce the synthetic spectra. So far, simulations have been done using a line by-line moderate resolution code, with a partially obsolete spectral lines database. We plan to develop an instrument-dedicated radiative transfer code, providing a parametric and analytic representation of monochromatic optical depths. Such a model, based on similar models already released for the Earth atmosphere (Amato et al. 2002), could help us also because it allows to calculate analytical jacobians, and to have a better control of the noise, that we have not included so far.

This is the first step to retrieve planetary atmospheric and surface parameters with the fully analytical and physical scheme described in Carissimo et al. (2005), Masiello et al. (2009) and Masiello & Serio (2013).

References

Amato, U., Masiello, G., et al. 2002, Env. Model. & Software, 17, 651
Bandfield, J. L., & Smith, M. D. 2003, Icarus, 161, 47
Carissimo, A., et al. 2005, Env. Modeling & Software, 20/9, 1111
Christensen, P. R., Bandfield, J. L., et al. 2001, J. Geophys. Res., 106, 23823
Conrath, B. J., et al. 1998, Icarus, 135, 501
Conrath, B. J., et al. 2000, J. Geophys. Res., 105(E4), 9509
Fonti, S., & Marzo, G. A. 2010, A&A, 512, A51
Formisano, V., et al. 2004, Science, 306, 1758
Heavens, N. G., et al. 2011, J. Geophys. Res., 116, E04003
Marzo, G. A., et al. 2006, J. Geophys. Res., 111, E03002
Masiello, G., et al. 2009, Atm. Chem. and Physics, 9, 8771
Masiello, G., & Serio, C. 2013, Applied Optics, 52(11), 2428
Mumma, M. J., et al. 2009, Science, 323, 1041
Rothman, L. S., et al. 1998, J. Quant. Spectrosc. Radiat. Transfer, 60(5), 665
Zahnle, K., et al. 2011, Icarus, 212, 493