Retrieval Advances of BrO/SO2 Molar Ratios from NOVAC

This Master thesis has been carried out by Elsa Wilken

at the

Institute for Environmental Physics, University of Heidelberg,

Germany

under the supervision of

Prof. Ulrich Platt

Optimierte Bestimmung des Molaren BrO/SO_2 Verhältnisses aus NOVAC Daten

Die Messung der absoluten Menge und des Verhältnisses von Vulkanischen gas Emissionen geben Einsicht in magmatische Prozesse. Das Network for Observation of Volcanic and Atmospheric Change (NOVAC) verfügt über ein System an automatisierten UV-Spektrometern, welche die Gas Emissionen der Vulkane aufzeichnen. Der Ausstoß von BrO und SO_2 kann mithilfe von Differenzieller optischer Absorptionsspektroskopie (DOAS) aus den aufgenommen Spekren bestimmt werden wobei die optische Absorption in der Fahne mit einem Hintergrundspektrum verglichen wird. Dies setzt voraus, dass das Hintergrund Spektrum nicht durch Vulkanische Gase beeinträchtigt ist. Typischerweise wird das Hintergrund Spektrum für einen Scan ein Höhenwinkel gewählt welcher als außerhalb der Fahne liegend identifiziert wird. Es hat sich jedoch gezeigt, dass auch diese Spektren noch durch Vulkanische Emissionen verunreinigt sein können. Als alternative Referenzspektren könnten 1) ein theoretisches Solar Atlas Spektrum oder 2) ein nicht verunreinigtes referenz Spektrum des selben Messgeräts dienen. Option 1) hat den Nachteil einer verringerten Messgenauigkeit, da Instrumenteneffekte hier modelliert werden müssen. Option 2) setzt vorraus, dass das Referenzspektrum unter ähnlichen Wetter- und Strahlungsbedingungen aufgenommen wurde. Wir verwenden die erste Methode um Kontaminierung zu identifizieren und greifen für die Bestimmung der Gas Konzentration auf die zweite Methode zurück um eine hohe fit Qualität sicher zu stellen. Stellen unsere Methode für NOVAC Daten von den Vulkanen Tungurahua und Nevado Del Ruiz vor.

Retrieval advances of BrO/SO2 molar ratios from NOVAC:

Measurements of magnitude and composition of volcanic gas emissions allow insights in magmatic processes. Within the Network for Observation of Volcanic and Atmospheric Change(NOVAC) automatically scanning UV-spectrometers are monitoring gas emission at volcanoes. The emissions of BrO and SO2 can be retrieved from the recorded spectra by applying Differential Optical Absorption Spectroscopy(DOAS) and comparing the optical absorption of the volcanic plume to the background. Therefore, the background spectrum must not be affected by volcanic influence. Classically, the background spectrum is taken from the same scan but from an elevation angle which has been identified to be outside of the volcanic plume. However, experience shows those background spectra can still be contaminated by volcanic gases. Alternatively reference spectra can be derived from 1) a theoretical solar atlas spectrum or 2) a volcanic-gas-free reference spectrum recorded by the same instrument. 1) comes with a drawback of reduced precision, as the instrumental effects have to be modeled and added to the retrieval. For 2), the alternative reference spectrum should be recorded at similar conditions with respect to meteorology and radiation. We use the first option to check for contamination and the second to evaluate the spectra to maintain a god fit quality. We present our approach and its results when applied on NOVAC data from Tungurahua and Nevado Del Ruiz.

Contents

7 Contamination based method

Based on the findings about the influence of external parameters on the BrO error we propose an algorithm which is able to pick an appropriate volcanic-trace-gas free reference.
The first step is, to evaluate every reference with solar atlas spectrum, to check for contamination.

If the reference is contaminated:

- We have a list of possible references where all references are not contaminated and the temporal distance to the plume date is no longer than 14 days.
- For all possible references we calculate the differences in the external parameters
- We use the analysis of external parameters as described above to estimate the BrO error of all references
- We choose the reference with the smallest estimated BrO error as new reference
- We evaluate the plume spectra with the new reference.

The assumption is, that the BrO error ϵ_{BrO} can be described as the sum of ϵ_0 and the deviation of ϵ_{BrO} with respect to all external parameters. ϵ_0 is the BrO error when evaluate the plume spectrum with the "same-time-reference". It is determined due to the accurateness of the NOVAC-instruments.

$$\epsilon_{BrO} = \epsilon_0 + \frac{d\epsilon}{dt} + \frac{d\epsilon}{d^\circ} + \frac{d\epsilon}{dT} + \frac{d\epsilon}{ddt} + \frac{d\epsilon}{dc} + \mathcal{O}(OP) \tag{7.1}$$

$$\rightarrow \Delta\epsilon_{BrO} = \epsilon_{BrO} - \epsilon_0 = \frac{d\epsilon}{dt} + \underbrace{\frac{d\epsilon}{d^\circ}}_{=0} + \frac{d\epsilon}{dT} + \frac{d\epsilon}{ddt} + \frac{d\epsilon}{dc} + \mathcal{O}(OP) \tag{7.2}$$

Here the parameter t stands for the time between plume-time and reference-time. The parameter T is the difference in temperature. The parameters dt and c are the differences in the daytime and the colorindex. The term (OP) accounts for other excluded external parameters. The task occurring at this stage is to find the best representation for the deviations. An then find the reference which minimize $\Delta\epsilon_{BrO}$

The easiest way is to just calculate the BrO error of all possible references for every plume by using the DOASIS routine. If this method is used it is possible to

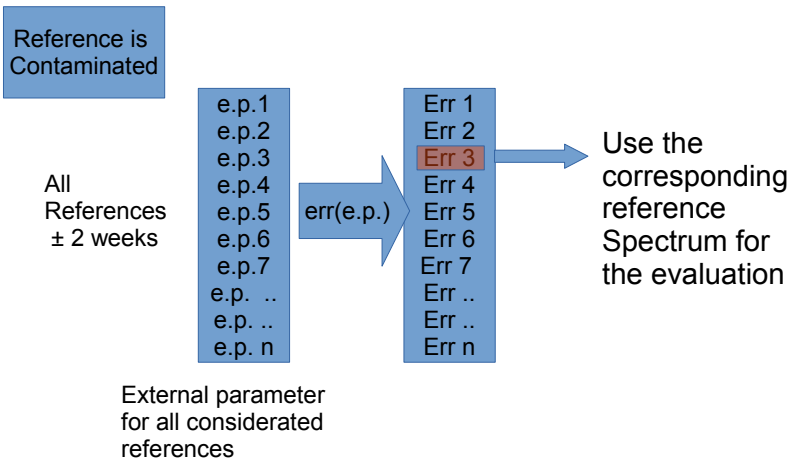


Figure 7.1: Visualization of the contamination based Method. If the reference is contaminated a list of possible references is available, where the temporal difference between the plume and the reference is not longer than two weeks. For every possible reference, the estimated BrO error is calculated by considering the corresponding difference in important external parameters. The reference with the so calculated minima BrO error is used for the evaluation.

choose the reference for which the BrO error is minimal. However this takes to much computation time since the evaluation time would be proportional to the number of possible references because the evaluation needs to be done for $\mathcal{O}(n)$ with n as the number of potential reference spectra. Doing this evaluation for every plume-reference pair makes it impossible to do the evaluation in real, or near real time.
In this thesis a novel approach of identifying an ideal reference spectrum, by considering externel parameters, is introduced. This way a much faster estimation with constant complexity $\mathcal{O}(1)$ is reached. But the above described optimal evaluation is used to rate new approach and compare them among each other. The optimal evaluation always choose the reference with the smallest absolute error. We don't use the relative error due to his vulnerability. Using the relative error leads to a less precision since the references with the highest BrO column density is preferred.
The results of the algorithm which chooses the reference automatically are described relative to an optimal evaluation. If the relative error is larger than 5 the data are excluded from the evaluation.

In the following we examine several methods for choosing the best reference based on the analysis of external parameters.

7.1 Fit data

The following chapter analyses fitting the data with a first order polynomial. Figure 6.6 to Figure 6.12 show the BrO error as a function of external parameters. Hereby, the curves are symmetric around zero difference in the respective external parameter. Therefore, it is not necessary to distinguish between positive or negative deviations from the equal surrounding conditions. Thus, the absolute differences can be utilized.

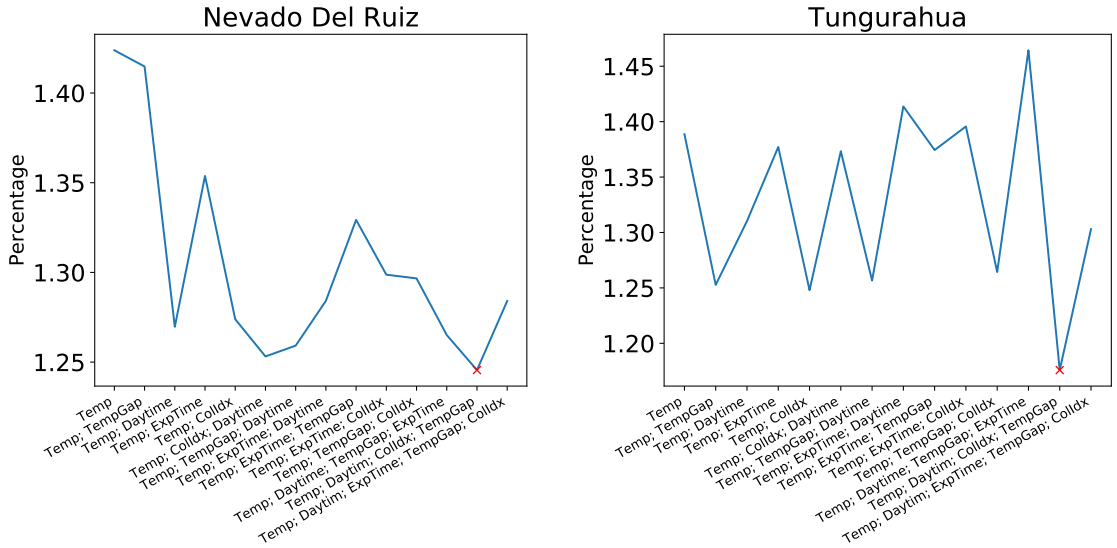
A linear approximation of the BrO error as function of the considered external parameters leads to a variation of Equation (7.2) : With linear differentiations of the BrO error with respect to the respective external parameters Equation (7.2) can be written as:

$$\Delta\epsilon_{BrO} = a_t \cdot \Delta t + a_{ET} \cdot \Delta ET + a_T \cdot \Delta T + a_{dt} \cdot \Delta dt + a_c \cdot \Delta c + \mathcal{O}(OP) \quad (7.3)$$

To determine the coefficients a_x (Equation (7.3)) the data visualized in Figure 6.6-6.12 where used. The fitting is done with an ordinary least square linear regression. In particular we used the python function Linear Regression from the library sklearn (SKl).

As it can be seen in Section 6.1 the impact of the different external parameters change for every instrument depending on the location and the instrument themselves. While the BrO error does not show any dependence on some external parameters for some instruments, the error has very strong dependence on the same external parameter at an other instrument. An example is the correlation between BrO error and the difference in daytime of 0.6 for D2J2201_0 (Nevado Del Ruiz) and a correlation of 0.16 for I2J8546_0 at the Tungurahua volcano. To get a more stable algorithm less external parameter are preferable. Thus, we need to distinguish between the stability of the fit, which improves with less external parameters and quality of the fit, which improves with more external parameters. A preferable strategy is, to find a solution which is valid for all instruments. Moreover, utilizing less external parameters saves computation time. One possibility is to use all external parameters where the correlation is above a certain value. Since we want to get a selection valid for all instruments there are two possibilities: The first one is that we decide by using the mean correlation of all instruments. The second option is to use the highest correlation.

To answer this question quantitatively for the fitting routine we evaluated data of Tungurahua and Nevado Del Ruiz with different combinations of the external parameter described in Section 6.1. Since we could not observe any correlation between the BrO error and the elevation angle the external parameter elevation angle was neglected in this analysis. To rate the results for the single instruments (three at Tungurahua and two at Nevado Del Ruiz) the difference to the "NOVAC-evaluation" is used. Hereby the factor X, a quantity which describes the distinction



(a) Nevado Del Ruiz

(b) Tungurahua

Figure 7.2: Deviation from the NOVAC-evaluation as a function of the selection of differences in external parameters which are used for the evaluation. Here the utilized combination of external parameters are plotted on the x-axis and the deviation on the y-axis. The ideal combination (lowest deviation) is marked with a red cross.

	Mean - Correlation	Highest - Correlation
Temperature	0.798	0.92
Colorindex	0.3108	0.409
Exposure time	0.265	0.452
Elevation Angle	0.02	0.067
Daytime	0.395	0.631

Table 7.1: The correlation coefficients between the BrO measurement error and the different external parameters. As a correlation value both the average and the maximum correlation is given.

between the NOVAC-Method and the contamination based method, serves as a indicator:

$$X = \frac{1}{n} \sum_{k}^{n} \frac{EContBased_k}{Enovac_k} \tag{7.4}$$

n is the total amount of contaminated spectra, *Enovac* is the relative BrO error, in the NOVAC-evaluation, *EContBased* is the relative BrO error, in the contamination based-evaluation. Figure 7.2 shows the X factor for the Tungurahua and the Nevado Del Ruiz volcano. The y-axis shows the factors X averaged over all instrument at the volcanoes.
The factors X change from instrument to instrument. The results for every instrument can be seen in the appendix (Figure .1). The factors X range from 1.18 to 1.42.
Single references-plume pairs result in a very high BrO error, which has a very high impact on the calculation of the X factors. As it can be seen in Figure 7.2 for both volcanoes the X factor is minimal for the combination of the following external parameters:

- Temperature
- Daytime
- Colorindex
- Temporal Difference

For the final algorithm this combination of external parameters is used. The coefficients a_x are calculated for each instrument at Nevado Del Ruiz and Tungurahua. Furthermore, the coefficients a_x are calculated with the combined data from all instruments installed at one volcano. The results for the Nevado Del Ruiz volcano can be found in Section 7.1 and for the Tungurahua volcano in Section 7.1.

Constant	(b) Data of Nevado Del Riz D2J2201_0 value	import	(c) Data of Nevado Del Riz D2J2200_0 value	import	(d) Data of Nevado Del Both Instruments value	import
a_T	7.505e+12	0.866	1.179e+13	0.925	1.07e+13	0.973
a_{CI}	3.48[illegible]13	0.048	1.024e+14	0.046	3.48e+10	0.070
a_t	-3.2e+09	0.0	-1.6e+09	0.0	-9.1e+08	0.0
a_{dt}	1.869e+12	0.095	1.054e+12	0.033	1.52e+11	0.006
					-6.81e+13	-0.047

Table 7.2: The results of the fitting with a first order polynomial. The constants of Equation (7.3) are calculated. The value shows the actual number. The importance, referred to as import, indicates the relative impact on the evaluation. (a)Data from Nevado Del Ruiz from the D2J2201_0 instrument. (b)Data from Nevado Del Ruiz from the D2J2200_0 instrument. Data from Nevado Del Ruiz from both instrument.

As it can concluded from Figure 7.2 The best results are found if all parameters are used with a mean correlation above 0.3 (see Section 7.1).

7.2 Other approaches

Fitting is not the only possibility of finding the optimal reference out of the list of all possible references.
In the following two additionally possibilities are presented. Both are based on the findings in section 6.1.

7.2.1 Nearest neighbor approach

Beside linear regression also the nearest neighbor approach can be utilized to estimate the BrO error for the evaluation with a potential reference spectrum. The nearest neighbor search describes an optimization problem for a given point $m \in \mathbb{R}^n$ and a set $S \subset \mathbb{R}^n$:

$$\bar{s}(m) = \min_{s \in S} d(m, s) \tag{7.5}$$

Here $d(\cdot, \cdot)$ is a distance function that computes the dissimilarity between the two input arguments. Typical distance metrics are the L1 distance $d_{\mathrm{L1}}(m, s) = ||m - s||^2$ or the L2 (also euclidean) distance function $d_{\mathrm{L2}}(m, s) = ||m - s||$.

In many cases not only one nearest neighbor but a set M of k nearest neighbors is of interest. In this case the optimization problem of Eq. 7.5 must be modified to

(b) Data of Tungurahua D2J2140_0

Constant	value	import
a_T	3.732e+12	0.664
a_{CI}	6.982e+13	0.078
a_t	3.5e+10	0.2
a_{dt}	7.797e+11	0.069

(c) Data of Tungurahua I2J8548_0

value	import
6.459e+12	0.920
2.760e+13	0.066
1.2e+10	0.13
-6.049e+11	-0.055

(d) Data of Tungurahua I2J8546_0

value	import
3.939e+12	0.850
-1.521e+13	-0.042
2.0e+10	0.2
1.158e+11	0.017

(e) Data of Tungurahua I2J8546_0

value	import
5.055e+12	0.838
2.534e+13	0.057
1.8e+10	0.117
-1.109e+11	-0.012

Table 7.3: The results of the fitting with a first order polynomial. The value shows the actual number. The importance, referred to as import, indicates the relative impact on the evaluation. The constants of Equation (7.3) are calculated. (a)Data from Tungurahua from the instrument. (b)Data from Tungurahua from the instrument. (c) Data from Tungurahua averaged over all instrument. both instruments Tungurahua

$$\bar{S}_k(m) = \min_{S_k \subset S} \sum_{s \in S_k, m \in M} d(s, m) \tag{7.6}$$

In many cases the nearest neighbor search is used to estimate a target variable y_m for a feature vector $m \in \mathbb{R}^n$. This method assumes a given set feature vectors S for which the target variables y_S are known. Then the target variable y_m for a given m can be estimated by:

$$y_m = \frac{1}{k} \sum y_S \tag{7.7}$$

Advantages of the nearest neighbour approach

The main advantage of the nearest neighbour method is that there is no need for a pre-assumption of a fitting method. The fitting method assumes that the BrO

error depends linearly on the external parameter. As it can be seen in Figure 6.6-Figure 6.12 this does not need to be the case. Thus, the nearest neighbour method is able to approximate the BrO error as a function of the external data in a more variable way.

Disadvantages of the nearest neighbour approach

Compared to fitting the data, the nearest neighbour is much slower. The reason for this is, the vast amount of compassion operations that is necessary for each computation. Thus, the calculation time increases with the amount of leaning data.
To normalize the distances d, the normalization function needs to be chosen and therefore pre-assumptions about feature importance are required. The x factor increases by 80%, from 1.24 for the fitting method to 2.24 for the nearest neighbour method. Including the fact, that the results of the nearest neighbor method is worse compared to the fitting method, the disadvantages of the nearest neighbour method outweigh the advantages.

7.2.2 Iterative approach

The idea of the iterative method is, that the importance of the individual external parameters are very different, that means if we have the list of possible references, we take all references where the temperature difference is minimal, so we get a new, much smaller list of possible references. From this list we choose all references where the next external parameter for example the daytime is minimal and again get a new list. We proceed this way with the following external parameters. We experiment with the sequence of the parameters, to increase the success of the method. The final sequence is:

$$Temperature \bullet Daytime \bullet Colorindex \bullet Temporal\ Difference \bullet Exposure\ time$$

Advantages of the iterative approach

The advantages of the iterative approach is the simple calculation and thus the short calculation time. Since the temperature is by far the external impact with the largest impact on the evaluation it is reasonable to look at first at the temperature, and after that on the other parameters.
Another advantage is, that the external conditions of the resulting references, are in an well-defined framework, similar to the conditions of the plume. Therefore the systematic errors resulting from the effects on the BrO SCDs (see Figure 6.20) can be kept small.

Disadvantages of the iterative approach

The iterative approach leads to an rigid evaluation of the data. References, which have very similar conditions as the plume could be excluded of the evaluation due

to an large difference in one external parameter. The decision for the best suited reference is based on one by one parameter. Thus it is impossible to look at all parameters by the same time. This could lead to a less optimal evaluation when looking at the BrO error. As a result the iterative method has a worse X factor (X=1.73) compared to the nearest neighbour (X=2.24) and the fitting method (X=1.24).