Machine learning methods for Precipitable Water Vapor estimation by radiometric data in millimetre wavelength

Grigoriy Bubnov¹, Peter Zemlyanukha¹, Evgeniy Dombek², Vyacheslav Vdovin¹,³

¹Institute of applied physics RAS, Nizhniy Novgorod, Russia
²University of Nizhniy Novgorod, Nizhniy Novgorod, Russia
³Nizhny Novgorod State Technical University, Nizhniy Novgorod, Russia

Corresponding author: bubnov@ieee.org

Abstract. This work deals with the first try to calculate the amount of Precipitable Water Vapor (PWV) in atmosphere by using machine learning and AI methods. We use the detector voltages series measured by radiometric system “MIAP-2” as the initial data for machine learning. The radiometer MIAP-2 works by “atmospheric dip method” in 2mm and 3mm atmospheric transparency windows. We also have PWV data series collected by Water Vapor Radiometer and GNSS receiver for data validation. The best convergence results were demonstrated by the independent component analysis (ICA) method with coefficient of determination $R^2 = 0.53$ and artificial neural network method (ANN) with $R^2 = 0.8$. These methods allow to reduce the systematic errors due to direct PWV calculation from raw radiometric data avoiding unnecessary steps opacity calculation.

1. Introduction

Studying astroclimate of a location is essential, when one chooses the most appropriate site for radio astronomical observations or space telecommunications. The suitability of a site is determined by a number of days per year, during which astronomical observations and space communication sessions are possible. In the millimeter range, astroclimatic conditions largely depend on the absorption of such waves by the atmospheric water vapor (Precipitable Water Vapor, PWV) and oxygen. This absorption depends on the elevation of the site, local and global climatic peculiarities, and weather; as the result, it may vary greatly from season to season and from day to day. The local climate is often responsible for how dry and wet air masses interchange over time, and, consequently, how the integral atmospheric absorption (or optical depth, or opacity) varies. The statistical data on the opacity are the chief astroclimatic parameter for the purposes of radio astronomy and space communications in the millimeter wavelength.

Meteorological conditions determine the astroclimate values, thus, the meteorological models and satellite data are often used for astroclimate predictions. The local climate makes the main contribution to opacity and it’s the reason to measure astroclimate right on the site of interest. Finally, the satellite data and model calculations together with local measurements complement each other, and give the full information of astroclimate conditions. Machine learning methods are often used in meteorology, including PWV estimations [1][2].
Radiometric data and methods are also wide used in meteorology and climate estimation. [3] The atmospheric dip method used by us in radiometric measurements has its drawbacks. The method is based on flat-layer atmospheric model, which is not always true. Cloudiness and other inhomogeneities in the atmosphere lead to significant errors in measurements, therefore, it is necessary to improve the calculation method. The first attempts to update the methodology were associated with the use of smoothing filters and adaptive algorithms [4] and proved to be quite good [5]. In this work, we attempted to compare the classical, modified method of a flat-layered model of the atmosphere and machine learning methods.

2. Equipment and Atmospheric dip method drawback

Our task is a rough-copy 2 and 3-mm atmospheric opacity measuring. With this objective in mind, we have developed instrumentation [6] and refined our methods [4], [5] for investigation of atmospheric propagation of terahertz waves. The “tau-meter” (the above mentioned MIAP-2) which we were using, was developed by the IAP RAS as the most appropriate instrument for the astroclimate measurements in the THz band. The radiometer allows us to estimate an integral absorption by using the atmospheric dip method. The hardware includes a radiometric system comprising two self-contained radiometers operating in two different bands of 84-99 GHz (λ ~ 3 mm) and 132-148 GHz (λ ~ 2 mm), a rotary support, a control, and a firmware system.

The well-known atmospheric dip method is widely used in radiometric measurements. [7] The method is based on flat-layer assumption, allowing to approximate the dependence of the sky brightness temperature on elevation by an exponent function and opacity is the exponent index. But in real atmosphere the flat-layer assumption is not met under cloudy conditions and inhomogeneities in the atmosphere. In fact, we have the collapsed voltage shift between the records, so we are unable to differentiate it against the background of noise. In this case the exponent approximation appears with a huge error. Figure 1 shows an example of raw-record of the “MIAP-2” radiometer detector voltages on different elevation angles. The atmosphere is close to flat-layered model in the clear weather conditions (left) and opposite in cloudy weather (right). The detector voltage is in direct ratio to brightness temperature on the corresponding angle and has a negative sign only for technical reasons. In classical atmospheric dip method these values are approximated by exponent, which leads to catastrophic errors of approximation in cloudy conditions.

![MIAP-2 raw record](image)

Figure 1 – The sky brightness record on different elevations was made on “Badary” observatory in January 2017. On the left is a clear sky, the atmosphere state is close to a flat-layered model. On the right - cloudy weather, unstable atmosphere.
3. A new data processing method

We are presenting a new approach for estimation of PWV by using atmospheric dip method, but without the flat-layer model. We use machine learning to produce the regression model between our data series and the PWV by Individual Component Analysis, (ICA) and k-Nearest Neighbours, (kNN) pipeline. We note that it is an only overview of an approach for the astroclimate study than the strict method. The essence of this approach is that we tried to estimate the amount of PWV without model assumption, but using independent measurements on another, more accurate device for training a neural network. The PWV data on Badary observatory is collected by two special meteorological instruments, which are well calibrated to PWV measurements and between themselves. Water Vapor Radiometer (WVR) measures the PWV at the Zenith from the sky brightness temperature at the slope of 22 GHz water line. [8] Global Navigating Satellite System receiver (GNSS) allows to calculate PWV from Zenith Path Delay by using time signals from navigation satellites. [9]. We also use the previous measurements made by atmospheric dip method in 2mm and 3mm atmospheric transparency windows on Badary observatory. Obviously, the tracks we observe carry a set of different mixed signals, wherein the contribution of water vapor absorption in the atmosphere must have nonzero covariance for the time series of detector voltages. Assuming this, we propose to use the dimension reduction procedure, which will make it possible to pass from the basis of the detector voltages to some new one, the components of which sufficiently represent the initial sample. A review of methods for choosing sufficiency is presented in the work [10]. After dimension reduction procedure, we estimate the degree of linear correlation of an individual component with a physical value – PWV by WVR and GNSS in our case.

There is a number of a dimension reduction procedure exists. To find the best one we compare three different methods: Principal Component Analysis (PCA), Individual Component Analysis (ICA) and Factor Analysis (FA) utilizing the scikit-learn (Scikit-learn: Machine Learning in Python [11]) by choosing the correlated with PVR. Since the correlation was far from linear, we use artificial neural network (ANN) and kNN as regression model and compare the results. The ANN has a hidden layer with 968 neurons with sigmoid activation function from Tensor Flow [12]. For the kNN we use 5 neighbors and m=2 metric. The detailed comparison and the analysis of the individual methods will be discussed in future papers. We use balanced ensemble divided to 2/3 training and 1/3 validation dataset.

Figure 2 – PWV calculating scheme from the raw-data of MIAP-2 by the classical method (above) and by the machine learning methods (below).

4. Data sets

For the method test we took the data obtained on Badary observatory from June 2016 to June 2017. [13] Observatory is located in sharp-continental climate of Sayan mountains, which means that opacity statistics should be low. However, due to cloudiness accumulating in the Tunkinskaya valley, the measurements were not the most successful. Relatively low values of integral humidity are evidenced by the readings of the meteorological equipment of the observatory. That was the reason to research the methods to minimize the influence of cloudiness on the results of measurements by the atmospheric dip method.
We have three rows of data at our disposal:

- **MIAP-2 data.** It is a radiometric system operating by atmospheric dip method in atmospheric transparency windows with wavelengths of about 3mm and 2mm. In fact, the system provides a raw records of brightness temperature of the sky on 6 elevations every 10 minutes. Thus, we have 12 data series: 6 for every channel.
- **Water Vapor Radiometer data (WVR).** It calculates an amount of PWV in Zenith by the intrinsic radiation of the atmosphere at the slope of the water line 22 GHz. [8]
- **GNSS receiver data.** This is a standard tool that allows, among other things, to determine the PWV from the Zenith Path Delay by using precise time signals from navigation satellites. [9]

The time resolution of these data sets is about the same, and it provides about 50 thousand points through the year. There are clear stable atmosphere weather and cloudy weather as well. The different weather in the sample allows us to train the processing algorithm in the best way.

5. Test sample calculation results

5.1. Dimension reduction

The greatest R²-factor (coefficient of determination of the prediction [14]) of the estimations the PWV by GNSS was obtained by ICA – 0.53 (by using the best component), FA – 0.39, PCA – 0.22. This result is a good argument that the time series we observe can be represented by independent mixed signals to a certain extent. These signals, in turn, are connected to the opacity and, probably, to clouds and other features of the real atmosphere. However, the nature of the dependence of the most correlated independent component with the PWV is nonlinear.

![Figure 3](image-url)

Figure 3 – The PWV regression between MIAP-2 data processed by ICA method and GNSS (left) and WVR (right). Blue dots indicate all data, and color are for predicted values.

5.2. Regression analysis

The picture below shows the data regression of PWV by ANN method with comparison to GNSS and WVR data. The WVR data was not involved in training and validation and provide a good comparison of our PWV estimation. There is no bias to the WVR data, both for the kNN and ANN regression. R² - factor of prediction was 0.86 for kNN and 0.8 for ANN. The kNN prediction is more linear, and the ANN are not, but we note that ANN can be better tuned. There is still the deviation from the trend line indicating presence of the stochastic noises. This deviation is much less important since we do the averaging in month bins for astroclimat analysis. But even with such data scatter, the resulting errors of PWV estimation by ANN and ICA methods are less then by old algorithm by opacity calculation. We note that the R² is lower if the dimension reduction procedure is excluded for kNN and ANN both.
The PWV are in a wide range of values, which suggests the possibility of using regression dependence and transformation to reduce the dimension for similar observations for other similar sessions. Finally, the methodic allows to determine the PWV bypassing the opacity calculation step and reducing systematic errors connected with uncertainty factors of specific absorption coefficients.

Figure 4 - The PWV regression between MIAP-2 data processed by ANN method and GNSS (left) and WVR (right). Blue dots indicate all data, and color are for predicted values.

6. Conclusion
In previous methods we used a MIAP-2 radiometric data to calculate opacity first, and then PWV values were calculated from opacity. At the first stage, errors related to the fitting of the exponent into the experimental data occurred and reached 40% in the worst conditions. [4] The calculation of PWV from opacity is connected with use of specific absorption coefficients, which can vary up to 300% depending on the site altitude, climate, calculation method and the model. [15] We offer a new algorithm which characterized by the fact that they don’t need to use specific absorption coefficients and bypass opacity calculation step. This reduces the systematic errors associated with the uncertainty of the specific absorption coefficients in water vapor and oxygen.

The next step will be validation the method on other data series, obtained in different places. We have 9-year experience of astroclimate measurements in different places of Eastern Hemisphere, so the problem of PWV estimation is quite actual for our research.

7. Acknowledgments
Astroclimate measurements are supported by the Russian Science Foundation under grant №19-19-00499. Data processing was carried out by Institute of Applied Physics RAS (theme 0030-2021-0005).

References
[1] D. Diouf, A. Niang, and S. Thiria, “DEEP LEARNING BASED MULTIPLE REGRESSION TO PREDICT TOTAL COLUMN WATER VAPOR (TCWV) FROM PHYSICAL PARAMETERS IN WEST AFRICA BY USING KERAS LIBRARY,” Int. J. Data Min. Knowl. Manag. Process, vol. 9, no. 6, 2019, doi: 10.5121/ijdkp.2019.9602.
[2] V. S. Kongara and D. Punyasesudu, “Forecasting the Water Vapour Distribution Over India Using Artificial Neural Networks,” no. July, pp. 1–12, 2015.
[3] A. F. Almansa et al., “Column Integrated Water Vapor and Aerosol Load Characterization with the New ZEN-R52 Radiometer,” Remote Sens., vol. 12, no. 9, p. 1424, Apr. 2020, doi: 10.3390/rs12091424.
[4] G. M. Bubnov, V. F. Vdovin, P. M. Zemlyanukha, and I. I. Zinchenko, “A refined method of the data processing for astroclimate measurements in mm-waves,” in 2018 29th IEEE International Symposium on Space Terahertz Technology, ISSTT 2018, 2018, pp. 164–168.

[5] G. Bubnov, V. Vdovin, P. Zemlyanukha, and I. Zinchenko, “Observations of astro-climate with the broad band radiometer using the atmospheric dip method,” in AIP Conference Proceedings, Dec. 2020, vol. 2300, no. 1, p. 020014, doi: 10.1063/5.0031757.

[6] V. I. Nosov et al., “A dual-wave atmosphere transparency radiometer of the millimeter wave range,” Instruments Exp. Tech., vol. 59, no. 3, pp. 374–380, May 2016, doi: 10.1134/S0020441216020111.

[7] A. G. Kislyakov, “On measuring the absorption of radio waves in the atmosphere by its thermal radiation” Radiotekh Eґlektron., 13, No. 7, 1161 (1968).

[8] Ilyin G.N., Troitsky A.V., "Determining the tropospheric delay of a radio signal by the radiometric method" Radiophysics and Quantum Electronics. vol. 60, no. 4, pp. 326–335, 2017, [Online]. Available: https://elibrary.ru/item.asp?id=29220875.

[9] Q. Zhao, Y. Yao, W. Yao, and S. Zhang, “GNSS-derived PWV and comparison with radiosonde and ECMWF ERA-Interim data over mainland China,” J. Atmos. Solar-Terrestrial Phys., vol. 182, no. June 2018, pp. 85–92, 2019, doi: 10.1016/j.jastp.2018.11.004.

[10] R. Cangelosi and A. Goriely, “Component retention in principal component analysis with application to cDNA microarray data,” Biol. Direct, vol. 2, no. 1, p. 2, Jan. 2007, doi: 10.1186/1745-6150-2-2.

[11] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

[12] M. Abadi et al., “TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems,” Mar. 2016, Accessed: Apr. 07, 2021. [Online]. Available: http://arxiv.org/abs/1603.04467.

[13] G. M. Bubnov, V. F. Vdovin, V. Y. Bukov, T. A. Makarov, G. N. Il’In, and I. I. Zinchenko, “Millimeter-wave astroclimate investigations on badary observatory near Baikal lake,” 2017 32nd Gen. Assem. Sci. Symp. Int. Union Radio Sci. URSI GASS 2017, vol. 2017-Janua, no. August, pp. 1–4, 2017, doi: 10.23919/URSIGASS.2017.8105000.

[14] Altman, Naomi S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". The American Statistician. 46 (3): 175–185. doi:10.1080/00031305.1992.10475879. hdl:1813/31637.

[15] G. Bubnov, P. Tremblin, V. Vdovin, P. Baron, and V. Khaikin, “Analysis of variations in factors of specific absorption of sub-terahertz waves in the earth’s atmosphere,” Proc. 2020 7th All-Russian Microw. Conf. (RMC), to be Publ. 2020., vol. 0, no. 1, pp. 1–4, 2020.