Lappeenranta University of Technology

Department of Computational Engineering

Major in Intelligent Computing

LUT
Lappeenranta
University of Technology

BM20A6100 Advanced Data Analysis and Machine Learning

Clustering of Satellite Data

Denis Sedov	0458147	Denis.Sedov@student.lut.fi
Sinan Kaplan	0458134	Sinan.Kaplan@student.lut.fi

TABLE OF CONTENTS

1. Introduction

As the climate change [1] has been becoming a significant issue in the present world due to the global warming, there are several actions that are taken internationally, particularly via Kyoto Protocol [2]. The main reason behind the climate change is Greenhouse gas (GHG) emissions [2]. The gases that contribute to greenhouse effect include water vapor, carbon dioxide (CO_2), Methane, Nitrous oxide and Chlorofluorocarbons (CFCs). Among these gases, especially CO_2 plays an important role more than rest of the other gases on greenhouse effect [4].

The CO_2 concentration has been increasing in the industrialized places throughout the world. That makes these places polluted and uninhabitable day by day and increases the level of global warming, thereby the side effect of climate change. Thus, it would be better to monitor such places over the time, take actions accordingly and provide more detailed information on the pollution levels. For instance, CO_2, CO, and NO_2 levels are used to provide such information in [5]. Recently, the study by Janne et al [6] gives the observation of CO2 on the regions of eastern USA, central Europe, and East Asia by the data gathered from Orbiting Carbon Observatory-2 (OCO-2) [7]. The proposed method is able to detect anthropogenic CO_2 emission areas by combining the information about CO_2 and NO_2 via clustering.

This work is an extension of Janne's work and it focuses on clustering task indicated in the study. First of all, we aim to study the impact of the grid size on the clustering task. Secondly, we focus on answering the following questions:

- What techniques there are for clustering ?
- How many clusters there are in the dataset ?
- How does the imputation of the missing values influence on the clustering results ?

The report is organized as follows: Section 2 describes the methods used for clustering analysis, Section 3 gives detailed analysis of experiments with respect to the problem settings and discusses findings and limitations of the study, and Section 4 summarizes all the process and gives possible future directions in this problem.

2. Methods

The problem defined above is studied by applying following methods for clustering: K-means [8], Expectation Maximization [9] and Gaussian Mixture Models [10]. The general steps in the proposed solution are illustrated in Figure 1. In the following subsections, each algorithm is explained in details.

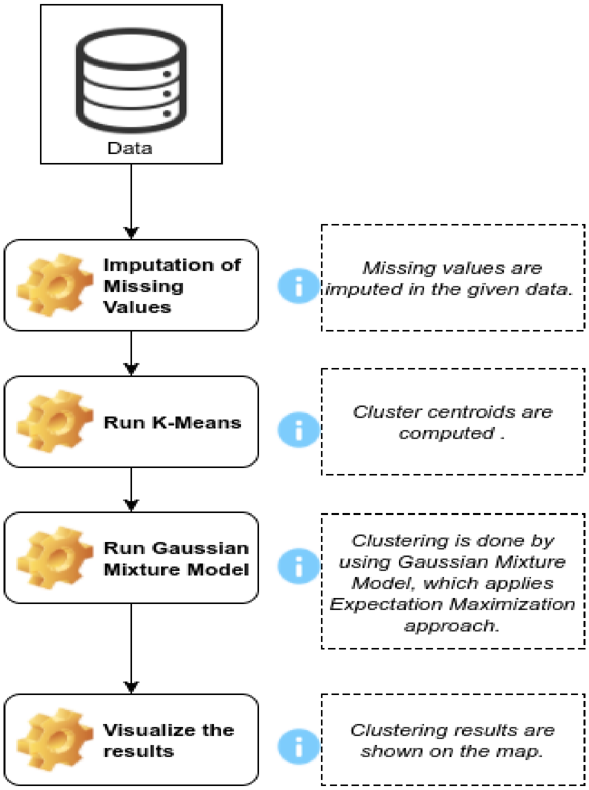
Data
Imputation of Missing Values
Missing values are imputed in the given data.
Run K-Means
Cluster centroids are computed .
Run Gaussian Mixture Model
Clustering is done by using Gaussian Mixture Model, which applies Expectation Maximization approach.
Visualize the results
Clustering results are shown on the map.

Figure 1. The flowchart of the proposed solution.

2.1. K-Means

One of the most known clustering algorithm is K-means [8], which is applied in this study to make an initial guess regarding the clusters in the data. The algorithm is quite easy to understand and apply to some of the clustering tasks.

To illustrate this approach, let $X = x\{i\},\ i = 1, ..., n$ be the set of n dimensional points to be clustered into K clusters, $C = \{c_k,\ k = 1, ..., K\}$. K-means tries to find groups in which squared error between the cluster center (mean) and the point is minimized. Let μ_k be the mean of cluster c_k and the cost function can be defined as follows:

$$J(c_k) = \sum_{x_i \varepsilon c_k} \left\| x_i - \mu_k \right\|^2 . \qquad (1)$$

As the goal of K-means is to minimize Eq. 1 over all clusters, this equation can be formulated as

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \varepsilon c_k} \left\| x_i - \mu_k \right\|^2 . \qquad (2)$$

The algorithm consists of following steps:

1. Initialize the clusters randomly.
2. Assign each point in the dataset to its closest cluster.
3. Compute cluster centers.
4. Repeat step 2 and 3 until convergence.

2.2. Expectation Maximization

In this study, Expectation maximization (EM) [9] is used for two tasks. The first one is to impute the missing values in the given dataset and the second one is for clustering as a part of Gaussian Mixture Model.

2.2.1. Imputation of Missing Values with EM

The given dataset contains many missing values that might have some effects on the clustering assignments. To reveal whether the clustering assignment is somehow affected by the presence of missing values, we first fill those missing values with EM by fitting a linear regression line. Expectation (E) simply imputes the missing values with respect to regression parameters and Maximization (M) updates the regression parameters using both imputed and complete data set.

To demonstrate this procedure, let $X_C = x\{i\},\ i = 1, ..., t$ be the complete set of X, $X_M = x\{i\},\ i = 1, ..., m$ be the set of X that contains missing values and θ be the regression parameters.

```
Input  :    $X_M$ and $X_C$
Step 1 :    Fit regression line to $X_C$ to estimate $\theta$
Step 2 :    While $TRUE$ (stopping criterion)
                E-Step : Fill $X_M$ using parameters $\theta$
                M-Step : Fit regression line to $X_M$ and $X_C$ to update $\theta$
```

2.2.2. Clustering with EM

The goal of EM clustering is to estimate the means and the standard deviations for each cluster so as to maximize the likelihood of the observed data (distribution). In other words, the EM algorithm attempts to approximate the observed distributions of values based on mixtures of different distributions in the clusters. EM approach for clustering is studied in the next section to illustrate Gaussian Mixture Models. However, the general steps are given below:

- E-step : each object is assigned to the centroid such that it is assigned to the most likely cluster.
- M-step : the cluster centroids are recomputed using least squares optimization.

2.3. Gaussian Mixture Models

Gaussian Mixture model (GMM) is one of the powerful techniques used for clustering [10]. The assumption in this model is that the data, $X = x\{i\},\ i\ =\ 1, ..., n$, is drawn from the density function,

$$f(x)\ =\ \sum_{k=1}^{K} p_k \Theta(x|\mu_k, \Sigma_k), \qquad (3)$$

where p_k is a mixture proportion $(0 < p_k < 1,\ k\ = 1, ..., K, \sum_{k=1}^{K} p_k = 1)$, $\Theta(x|\mu_k,\ \Sigma_k)$ stands for multivariate Gaussian distribution with mean μ_k and variance Σ_k. The mixture parameters $\theta\ =\ (p_{1,}, ..., p_k,\ \ \mu_1, ..., \mu_k,\ \Sigma_1, ..., \Sigma_k)$ are estimated by maximum likelihood as follows.

$$L(\theta|x_1, ..., x_n)\ =\ \sum_{i=1}^{n} \left[\sum_{k=1}^{K} p_k \Theta(x|\mu_k,\ \Sigma_k) \right]. \qquad (4)$$

To find the maximum likelihood, EM algorithms is applied. The detailed explanation of the method is described in the study [10]. In this context, GMM procedure can be illustrated using the steps below:

```
Input   Initialize parameters ($K$, $\mu_k$, $\Sigma_k$)

        While $TRUE$ (stopping criterion)

                E-Step : Compute conditional probability $\widehat{p}_k(x)$,

                M-Step : Update parameters $\theta\ =\ (p_{1,}, ..., p_k,\ \ \mu_1, ..., \mu_k,\ \Sigma_1, ..., \Sigma_k)$
```

3. Experimental Analysis and Discussion

This section describes the whole experimental analysis of methods both with toy data (in this case Mickey Mouse) and the given data. The given data consists of four different grid size (2×2, 1×1, 0.5×0.5, 0.25×0.25 degree by degree resolution) and it is used to study the impact of the grid size on the clustering. Experiments with the actual data are done both with imputation of missing values and without it. That is why this section is divided into following subsections: Section 3.1 gives the results obtained from toy data, Section 3.2 explains the results acquired from the original data without any preprocessing and Section 3.3 provides results with the imputed missing values in the given data. The results are discussed under each section. The whole implementation of the project with related figures and explanations can be seen in GitHub repository [11].

3.1. Experiments with Toy Data

K-means and GMM clustering algorithms are first run on the toy data to reveal whether they provide any significant results or not. The results of clustering are shown in Figure 2 and one can clearly see that GMM superiors K-Means on this data set. This indicates that clustering based only on the distance measurement is not always suitable for the specific datasets.

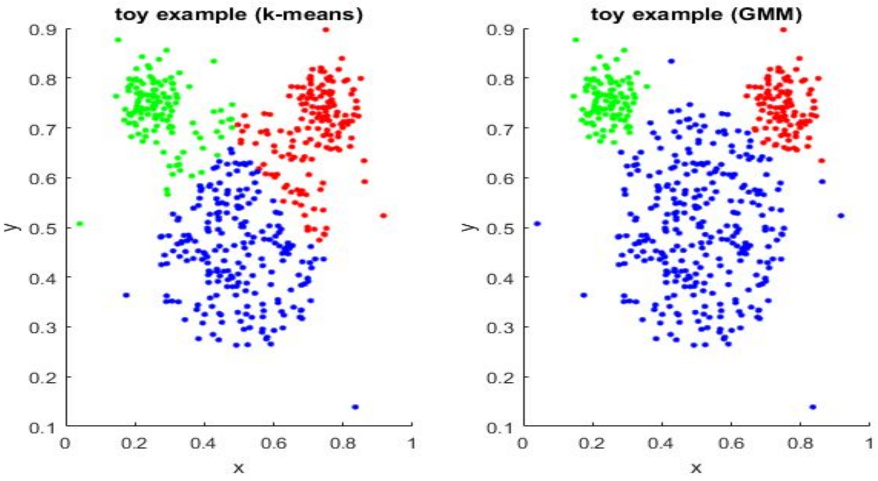

Figure 2. K-Means and GMM clustering on the mouse data.

3.2. Experiments without Imputation

First of all, we have run the GMM based clustering with ***k = 3, 4*** *and* ***5*** as the number of clusters to reveal the appropriate number of clusters to use further in this project. We have approached this problem both from engineering perspective and the number of mixture components. Here, we are interested in highly polluted areas such as Eastern Asia, Central Europe and the eastern part of USA, which emit more CO_2 than the rest of the areas on the map. Therefore, it is important to conduct the analysis in these regions by dividing them further into clusters in order to find the areas of different pollution level. When the number of clusters is small (e.g. k = 3), the clustering process gives only rough estimation of the polluted regions, however these regions could be further divided into subregions if the greater number of cluster were used and therefore it would give us more valuable information. On the other hand, when the number of clusters is higher than four (k=4), the polluted areas (marked with yellow and green color) are not divided and only the areas covered by water (like sea and lakes) are divided further and these particular areas doesn't carry any useful information regarding CO2 and NO2 emission. For instance, one can compare the European region when the number of cluster is three and four. From this area it can be seen that the map with k = 4 provides more detailed information than the map with k = 3. Afterwards, one can compare the same region when the number of cluster is four and five. In this case, it is obvious that this region is not divided any further and the patterns extracted from the region remain the same. Thus, it is not necessary to increase the number of clusters in order to extract the information about highly polluted areas, since it will not give any useful results. Clustering results for each particular grid size are shown in Appendices (see Appendix 1 for grid size 0.25x0x25, Appendix 2 for grid size 0.5x0.5, Appendix 3 for grid size 1x1 and Appendix 4 for grid size 2x2). The figures related to k = 3 and 5 can be seen on Github [11] repository under the relevant folders.

Replacement of Figure 3, Figure 4, Figure 5 and Figure 6 as an Appendix

The grid size plays an important role in the clustering analysis. When the grid size increases, one can see from the figures that the polluted areas (green and yellow regions) are represented more accurately, regarding the size and the shape of the region, which helps to identify the regions with the highest pollution more precisely. For instance, when we look at the USA map on the Figure 6 with the grid size 2x2, the yellow region encompasses the large area. However, the same region on the Figure 4 (with smaller grid size) is divided in 2 clusters, revealing some additional information about the places with the highest pollution. Thus, the large grid size gives only rough estimation of the polluted regions and some valuable information might be missing. In addition to that, due to the missing values, some important locations do not contribute to the clustering and this causes sparsity on the region when the grid size is small.

3.3. Experiments with Imputation of Missing Values

To address the effect of imputation on the given task in the case of missing values, we have implemented EM based imputation as mentioned in Section 2.2.1. The clustering is done on the data set with *2×2* and *1×1* grid size, the number of clusters k = 4. The results show that the imputation of missing values seems not a good choice for this project. One can see the clustering results in Figure 3. The main reason behind this situation is that some instances have missing values in both columns of the data (NO_2 and CO_2) and after the imputation these instances take same value for each column. That decreases the overall quality of the clustering due to the linearity introduced by imputation. For the grid size 0.5×0.5 and 0.25×0.25 it causes ill-conditioned variance problem during the clustering. One can think of removing the instances which have missing values on both columns and then perform the imputation. However, for the clarity of this project we have skipped this process and focused on the clustering analysis.

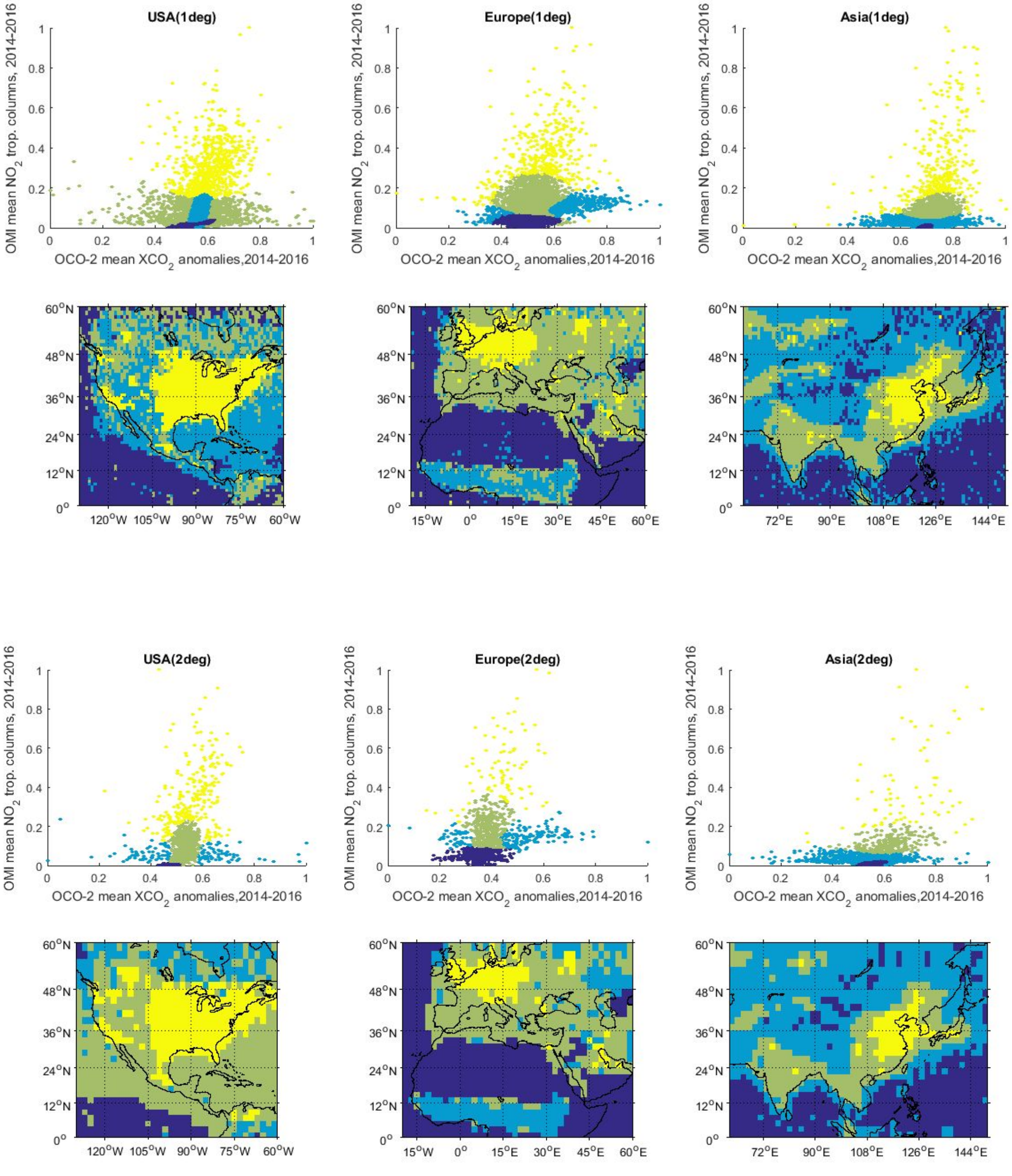

Figure 7.Clustering Analysis with imputation: (top) 1x1 grid size, (bottom) 2x2 grid size.

4. Conclusion

In this paper, the impact of the grid size on the clustering analysis is studied based on the paper [6]. Clustering analysis is done both with imputation of missing values and without it. The clustering analysis for each region without imputation is done based on the different grid sizes and it is stated that the optimal number of clusters in the given dataset is four (k = 4). It is found by analyzing CO2 concentration over the polluted regions of USA, Europe, and Asia. The obtained results show that when the grid size increase we can have more accurate information about the polluted regions(e.g. the shape and the size of the region). When the missing values are filled by EM approach, the obtained results suffer from ill-conditioned variance problem because of the linearity introduced by the imputation.

One can think of trying different clustering techniques, such as kernel K-Means, for further analysis of the given data. In addition to that, decision tree based clustering methods can be used to have country - to country comparison of CO2 pollution. Also, one can further try to estimate what causes CO2 pollution in each country over the season.

REFERENCES

[1] "NASA: Climate Change and Global Warming,". [Online]. Available: http://climate.nasa.gov. Accessed: Jan. 23, 2017.

[2] "Kyoto protocol," United Nations Framework Convention on Climate Change. [Online]. Available: http://unfccc.int/kyoto_protocol/items/2830.php. Accessed: Jan. 23, 2017.

[3] Naomi Oreskes, *"The Scientific Consensus on Climate Change,"* Science 3 December 2004: Vol. 306 no. 5702 p. 1686 DOI: 10.1126/science.1103618

[4] “Intergovernmental Panel on Climate Change (2013), Climate Change 2013: The Physical Science Basis”, 1535 pp., Cambridge Univ. Press, Cambridge, U. K., and New York.

[5] Lindenmaier, Rodica, et al. "Multiscale observations of CO2, 13CO2, and pollutants at Four Corners for emission verification and attribution." *Proceedings of the National Academy of Sciences* 111.23 (2014): 8386-8391.

[6] Hakkarainen, J., I. Ialongo, and J. Tamminen. "Direct space-based observations of anthropogenic CO2 emission areas from OCO-2." *Geophysical Research Letters* 43.21 (2016).

[7] Crisp, David, et al. "The orbiting carbon observatory (OCO) mission." *Advances in Space Research* 34.4 (2004): 700-709.

[8] Jain, Anil K. "Data clustering: 50 years beyond K-means." *Pattern recognition letters* 31.8 (2010): 651-666.

[9] Do, Chuong B., and Serafim Batzoglou. "What is the expectation maximization algorithm?." *Nature biotechnology* 26.8 (2008): 897.

[10] Biernacki, Christophe, Gilles Celeux, and Gérard Govaert. "Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models." *Computational Statistics & Data Analysis* 41.3 (2003): 561-575.

[11] Github Repository - Clustering of Satellite Data," GitHub, 2017. [Online]. Available: https://github.com/kaplansinan/Clustering-of-Satellite-Data . Accessed: Jan. 29, 2017.

APPENDICES

Appendix 1

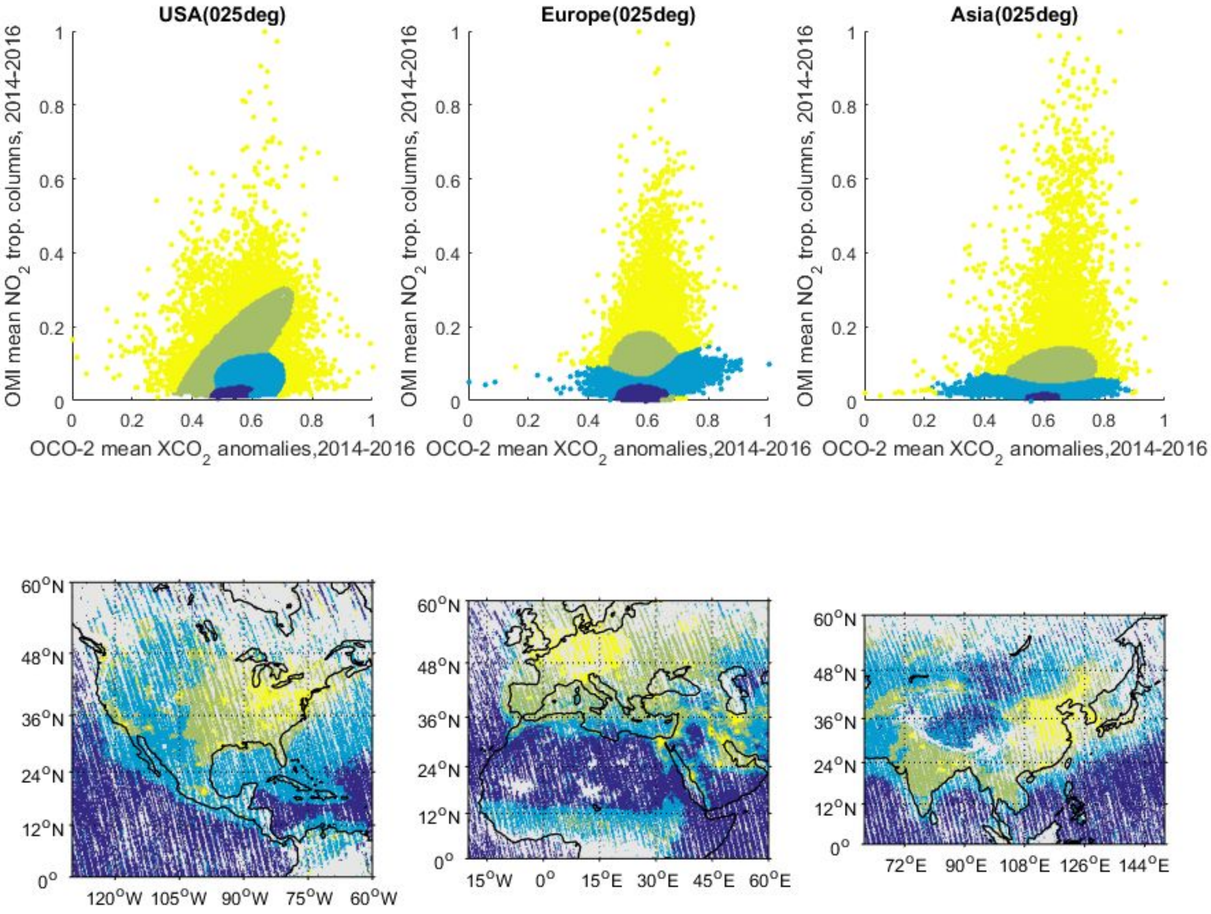

Figure 3. Clustering Analysis of Satellite data with GMM - grid size (0.25x0.25).

Appendix 2

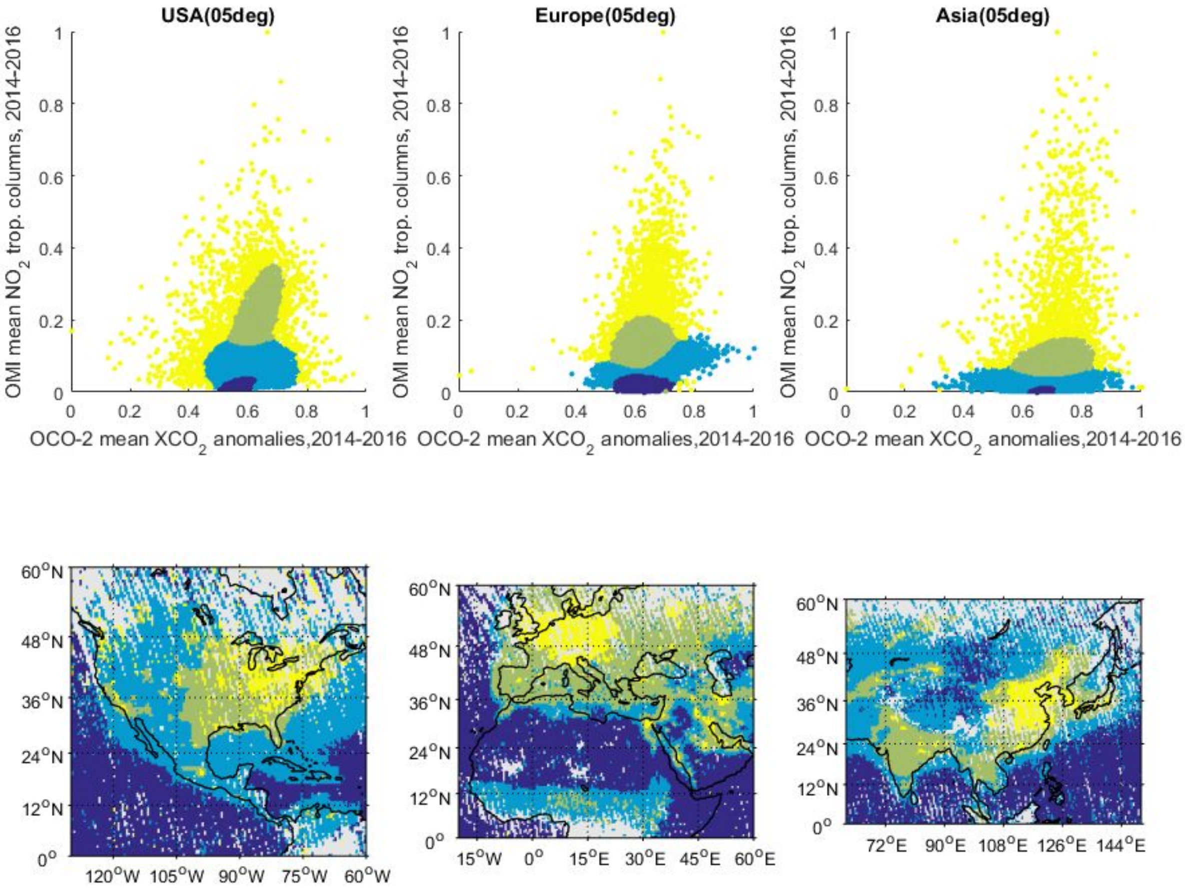

Figure 4. Clustering Analysis of Satellite data with GMM - grid size (0.5x0.5).

Appendix 3

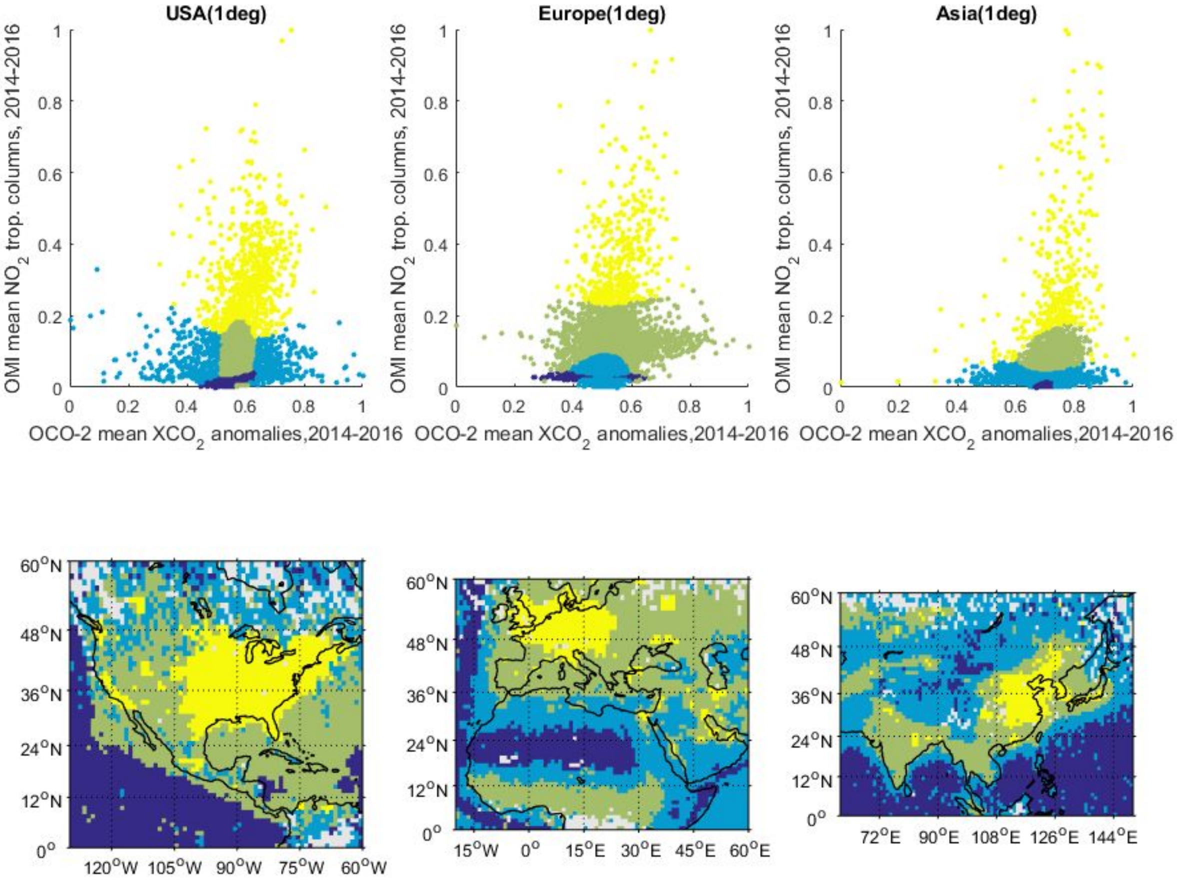

Figure 5. Clustering Analysis of Satellite data with GMM - grid size (1x1).

Appendix 4

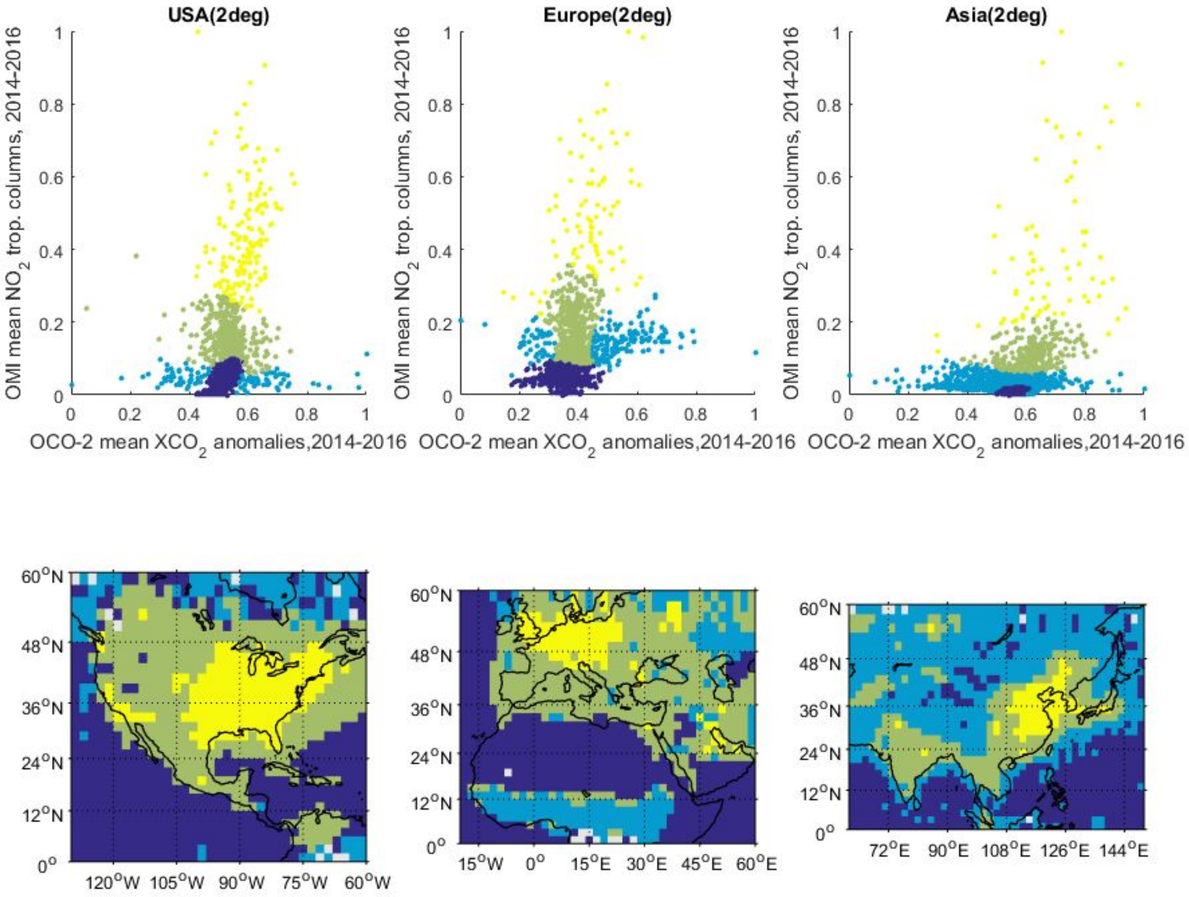

Figure 6. Clustering Analysis of Satellite data with GMM - grid size (2x2).