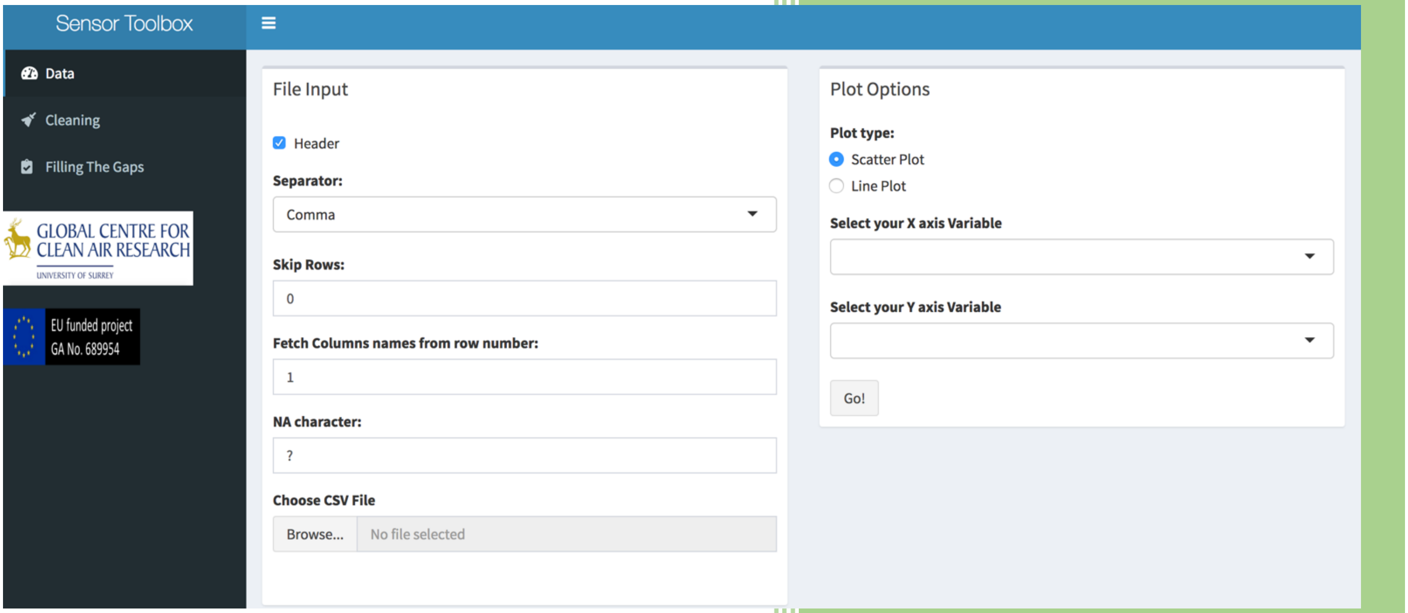
Sense Your Data: Sensor Toolbox Manual, Version 1.0



File Input

☑ Header

Separator: Comma

Skip Rows: 0

Fetch Columns names from row number: 1

NA character: ?

Choose CSV File: No file selected

Plot Options

Plot type:
● Scatter Plot
○ Line Plot

Select your X axis Variable

Select your Y axis Variable

Sachit Mahajan & Prashant Kumar*

Global Centre for Clean Air Research (GCARE), Department of Civil and Environmental Engineering, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford GU2 7XH, United Kingdom

Summary

The use of low-cost sensors for environmental monitoring has led to a significant increase in the data volume and availability, which has made data processing and its analysis a challenging task. Here, we present Sense Your Data: Sensor toolbox, a dashboard hosted on R Shiny platform that can assist researchers as well as people from non-technical background to analyse and visualise data in an easy way. The tool supports several functions like data summary, plotting, outlier detection and gap filling.

Introduction

Sense Your Data: Sensor toolbox is a web based data analysis toolbox that is designed for time-series data analysis and visualization using R Shiny (Ishimaru et al., 2014) platform. Primarily, it is developed for analysing and visualising air quality data. However, it can be applied to other type of data-sets as well. The motive behind having such a tool is to have an easy and efficient way to analyse data for researchers as well as non-researchers. An important design aspect was to have a simple dashboard which is easy to understand and operate, which enables a user to simply input the data file and analyse it. The graphical user interface reduces the task of writing or changing the R scripts to perform basic analysis of time-series data. Sense Your Data: Sensor toolbox should be easy to operate even for inexperienced users who want to visualize and analyze the data.

The main features of the Sensor Toolbox include:

- Analyzing and visualizing data without writing scripts
- Better understanding of data using adjustment of parameters.
- Better data visualization using scatter and line plots.
- Data analysis features like data summary, outlier detection and gap filling.

The application has been tested on Windows and Mac OS and works fine for different platforms.

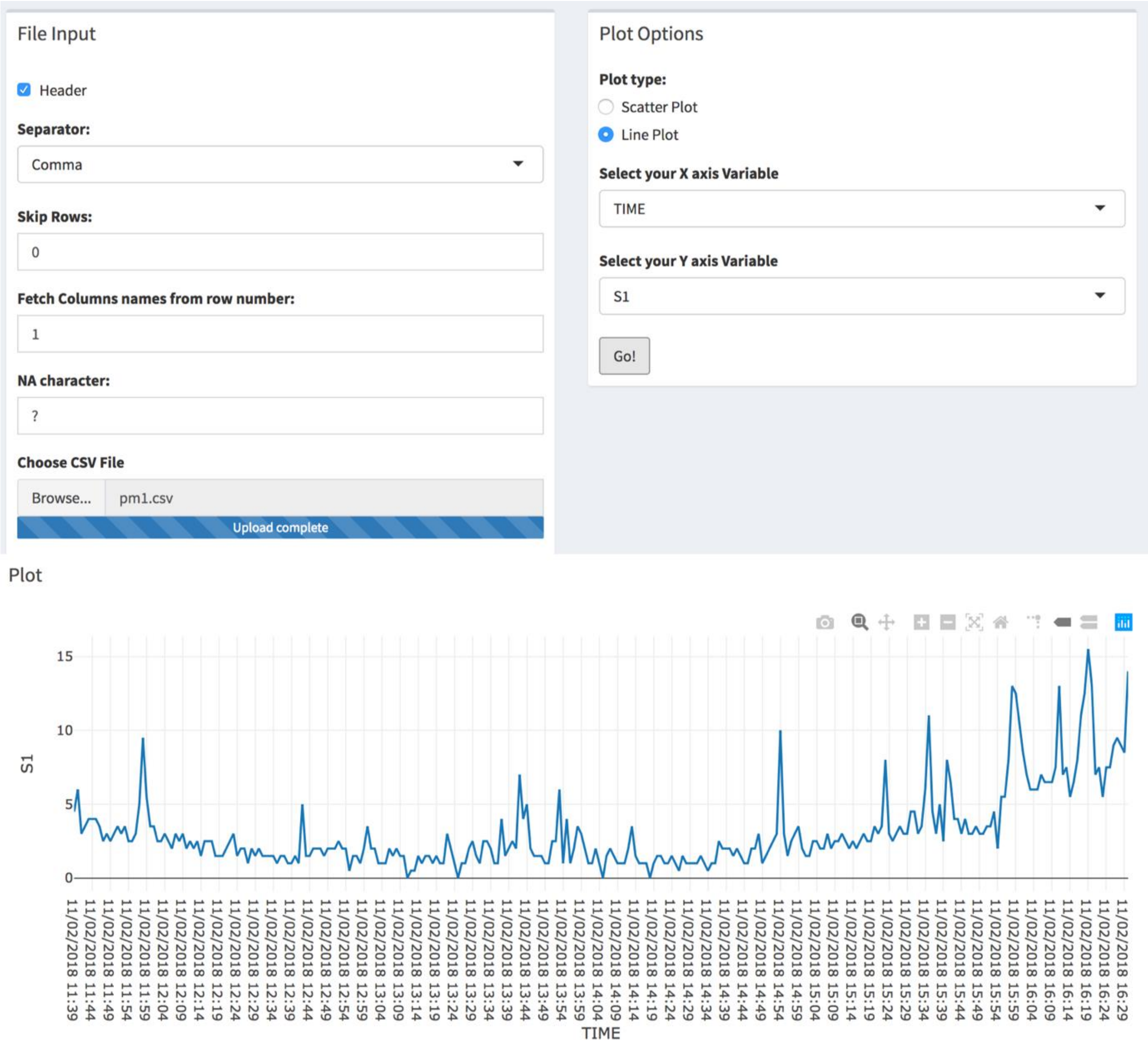
Functions

The Sensor Toolbox has several functions for data analysis and visualization. They are mainly divided into three tabs:

1. **Data** – This tab includes the file input option. The user can input any .csv file or with separator (comma, semicolon, tab). Once the file is uploaded, the user would be able to view the raw data as well as the data summary at the lower part of the dashboard. Based on the requirements, a user can skip rows, enable/disable headers as well as specify which column names to fetch and from which row. In addition to that the user can also specify what kind of missing data is there in the file. The user can write it down in the “NA character” box. It can include “NA”, “Nan”, “?”, “N/A”, “none”.

Citation for this document:
Mahajan, S., Kumar, P., 2019. Sense Your Data: Sensor Toolbox Manual, Version 1.0. pp.1-7. DOI: 10.13140/RG.2.2.17249.76640/4

Once the data is uploaded, there are option to plot the data using “Plot Options” on the right side of the dashboard. The user can have a line plot or a scatter plot by selecting the X and Y axis variable as shown in Figure 1.



File Input

☑ Header

Separator: Comma

Skip Rows: 0

Fetch Columns names from row number: 1

NA character: ?

Choose CSV File: Browse... pm1.csv

Upload complete

Plot Options

Plot type:

○ Scatter Plot

● Line Plot

Select your X axis Variable: TIME

Select your Y axis Variable: S1

Go!

Plot

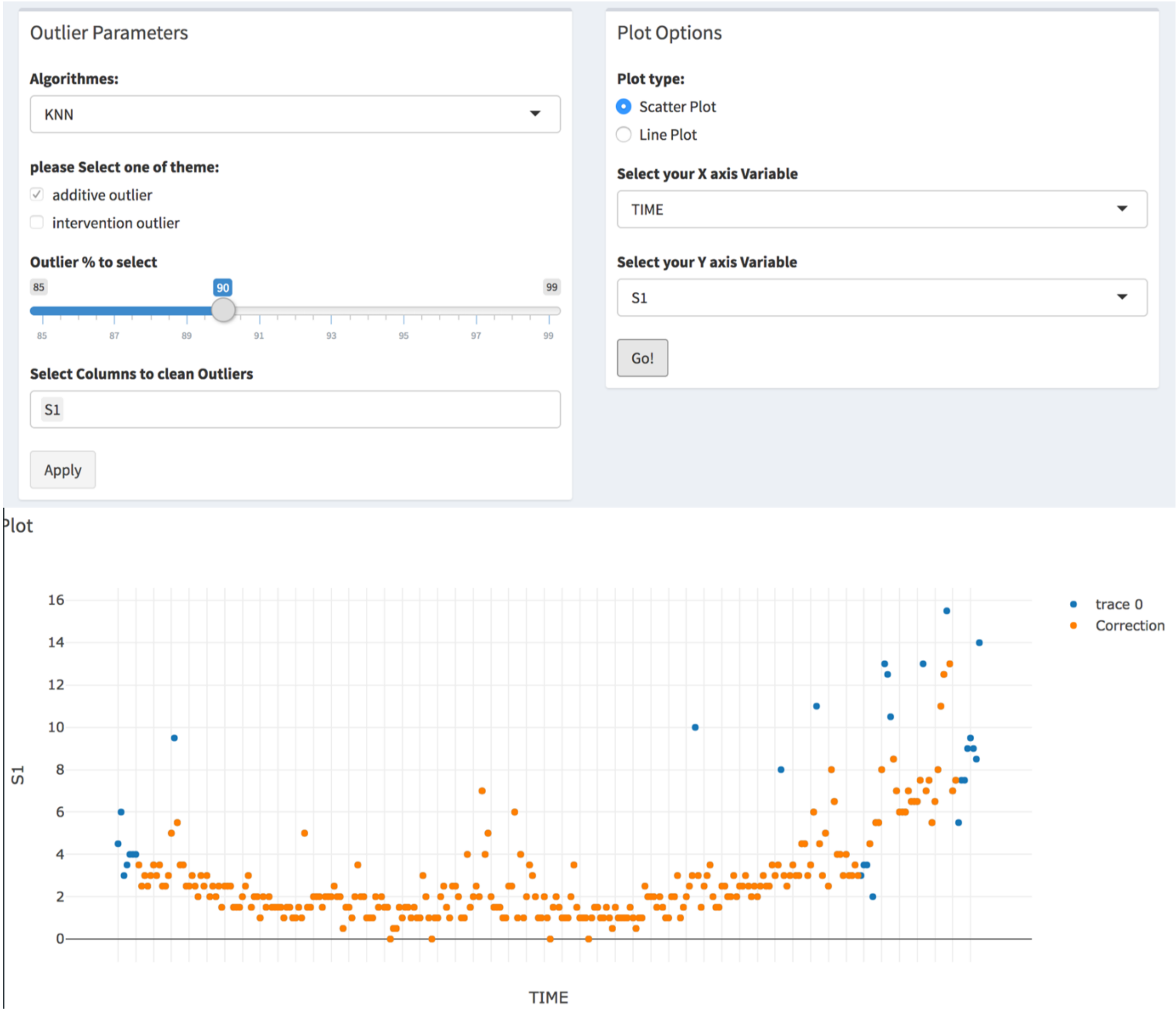
Figure1. Data upload and plotting

2. **Cleaning** – This tab basically addresses the challenge of cleaning the data by removing any outlier or anomaly. When dealing with a large amount of time-series data, it is important to make sure that the data is reliable and without outliers (Chen et al., 2018; Ottosen and Kumar, 2019). There are three different algorithms that have been implemented:
    - Autoregressive Integrated Moving Average (ARIMA) Additive : ARIMA model has been widely used for tasks related to prediction and forecasting (Mahajan et al., 2018a). An additive outlier means an islotaed spike in the time-series data. The model is implemented using “*tsoutliers*” package (López-de-Lacalle, 2016) in R.

Citation for this document:
Mahajan, S., Kumar, P., 2019. Sense Your Data: Sensor Toolbox Manual, Version 1.0. pp.1-7. DOI: 10.13140/RG.2.2.17249.76640/4

- K-nearest neighbor (KNN): In KNN anomaly detection method, each point in the dataset get an anomaly score which is the distance to its kth nearest neighbor in the data. Based on that, all the points are given a ranking that depends on the distance. If the score is higher, the chances of it being an anomaly are higher (Ramaswamy et al., 2000). While implementing this model, users can select the percentage of outliers they want using the slider bar option in the dashboard.
- Autoencoder: Artificial Neural Network (ANN) models have been widely used for time-series data for dealing with tasks like air quality forecast (Mahajan et al., 2018b). An autoencoder neural network follows an unsupervised learning algorithm that uses a backpropagation model to set the weights that attempts make the outputs to be equal to the input values. This model is implemented using the "autoencoder" (Eugene Dubossarsky and Tyshetskiy, 2015).

Once the user has selected the outlier detection algorithm, the next step involves selecting the columns to show the outliers and visualize them as shown in Figure 2. The data with the outliers can be plotted using the plot feature of the dashboard. The plot would show the "trace", that is the original value and "correction", that represents the value that needs to be corrected.



Outlier Parameters

Algorithmes: KNN

please Select one of theme:

☑ additive outlier

☐ intervention outlier

Outlier % to select: 90

Select Columns to clean Outliers: S1

Plot Options

Plot type:

● Scatter Plot

○ Line Plot

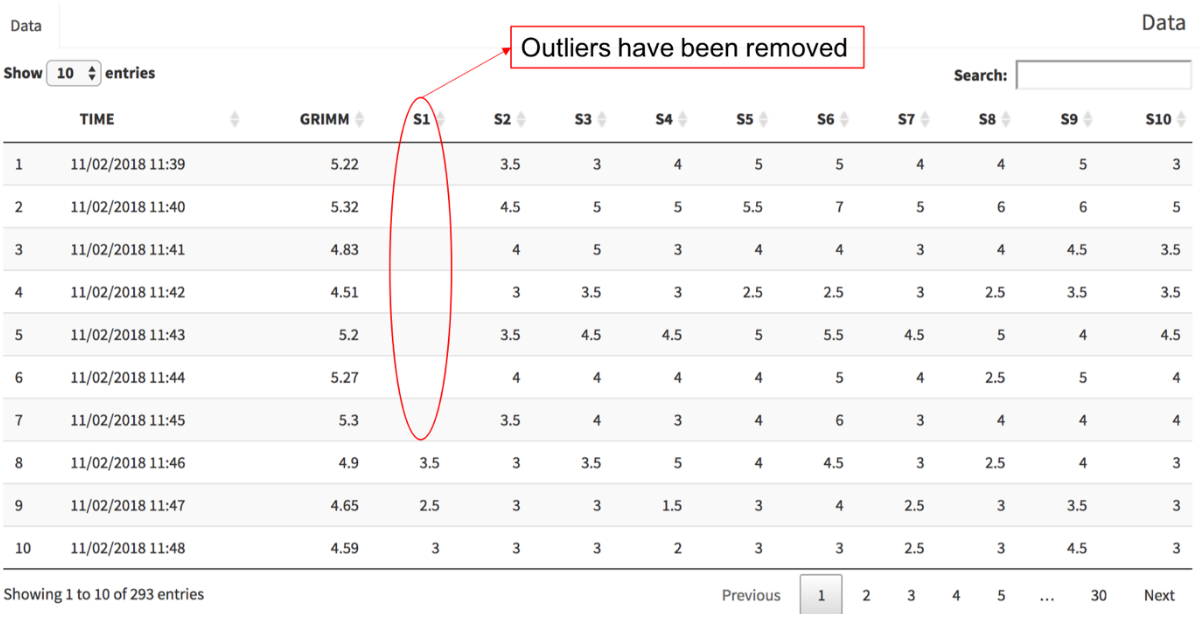
Select your X axis Variable: TIME

Select your Y axis Variable: S1

Figure 2. Application of outlier detection method and visualization of outliers

Figure 3 shows one example of for outlier detection function where the outlier values are removed from the dataset

Citation for this document:
Mahajan, S., Kumar, P., 2019. Sense Your Data: Sensor Toolbox Manual, Version 1.0. pp.1-7. DOI: 10.13140/RG.2.2.17249.76640/4



Outliers have been removed

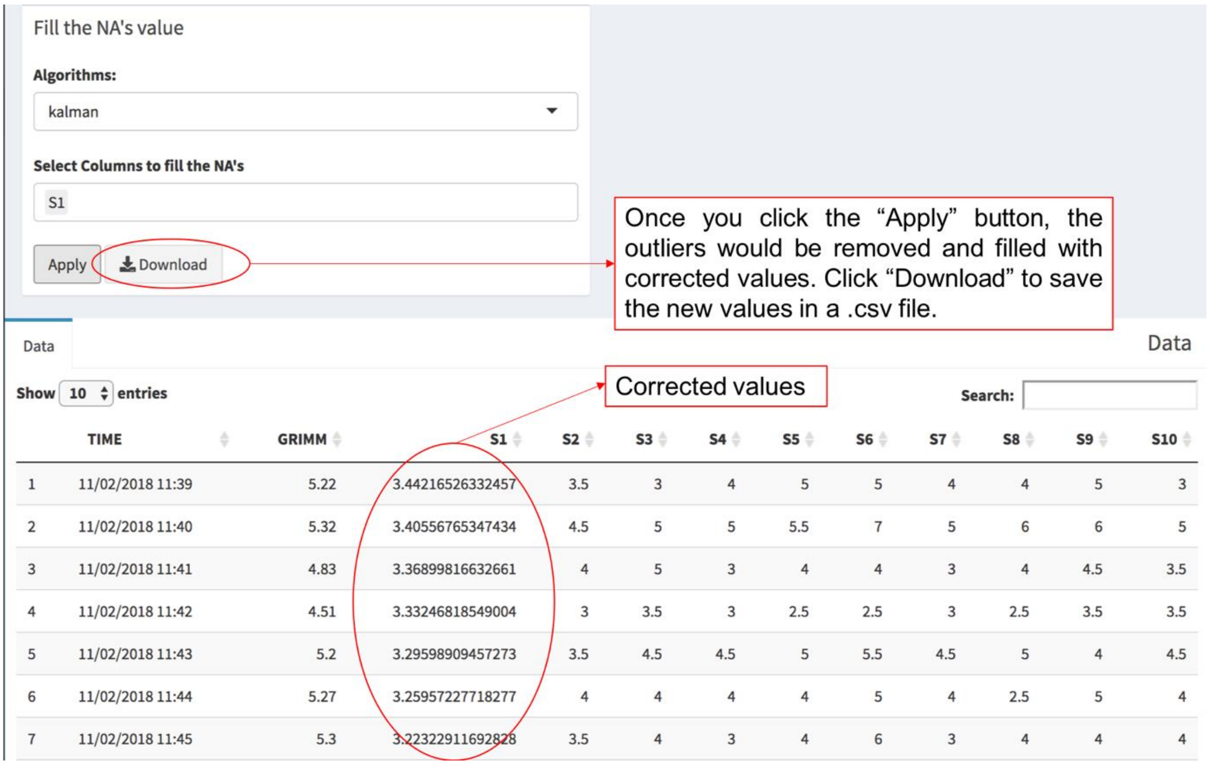
	TIME	GRIMM	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	11/02/2018 11:39	5.22		3.5	3	4	5	5	4	4	5	3
2	11/02/2018 11:40	5.32		4.5	5	5	5.5	7	5	6	6	5
3	11/02/2018 11:41	4.83		4	5	3	4	4	3	4	4.5	3.5
4	11/02/2018 11:42	4.51		3	3.5	3	2.5	2.5	3	2.5	3.5	3.5
5	11/02/2018 11:43	5.2		3.5	4.5	4.5	5	5.5	4.5	5	4	4.5
6	11/02/2018 11:44	5.27		4	4	4	4	5	4	2.5	5	4
7	11/02/2018 11:45	5.3		3.5	4	3	4	6	3	4	4	4
8	11/02/2018 11:46	4.9	3.5	3	3.5	5	4	4.5	3	2.5	4	3
9	11/02/2018 11:47	4.65	2.5	3	3	1.5	3	4	2.5	3	3.5	3
10	11/02/2018 11:48	4.59	3	3	3	2	3	3	2.5	3	4.5	3

Showing 1 to 10 of 293 entries

Figure 3. Outliers are removed from the dataset

3. **Filling the gaps** – This tab deals serves two purposes: a) gap filling and b) downloading the csv file after outlier correction. For both the features, the user would need to select the column name (it can be a column which has missing data or the column that has been cleaned from outliers in the previous tab). For gap filling, there are two algorithms which have been implemented:
    - Interpolation: In general terms, interpolation refers to a method of creating new data points within a set of known data points. The interpolation function using linear interpolation to fill the missing values. This is implemented using the “imputeTS” package (Moritz, 2018).
    - Kalman filter: The kalman filter represents a set of mathematical equations that provides recursive means to do an estimation of the state of the process in such a way that it minimizes the mean squared error. The Kalman filter can supports several features like past data estimation, present and future values. It can also do these tasks when the precise nature of the system is unknown (Welch and Bishop, 2006).

    Once the data is cleaned, the user can download the new data by specifying the column names in the option provided and click the download button. An example has been provided in Figure 4. The new file would be downloaded in the .csv format.



Fill the NA's value

Algorithms: kalman

Select Columns to fill the NA's: S1

Once you click the “Apply” button, the outliers would be removed and filled with corrected values. Click “Download” to save the new values in a .csv file.

Corrected values

	TIME	GRIMM	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	11/02/2018 11:39	5.22	3.44216526332457	3.5	3	4	5	5	4	4	5	3
2	11/02/2018 11:40	5.32	3.40556765347434	4.5	5	5	5.5	7	5	6	6	5
3	11/02/2018 11:41	4.83	3.36899816632661	4	5	3	4	4	3	4	4.5	3.5
4	11/02/2018 11:42	4.51	3.33246818549004	3	3.5	3	2.5	2.5	3	2.5	3.5	3.5
5	11/02/2018 11:43	5.2	3.29598909457273	3.5	4.5	4.5	5	5.5	4.5	5	4	4.5
6	11/02/2018 11:44	5.27	3.25957227718277	4	4	4	4	5	4	2.5	5	4
7	11/02/2018 11:45	5.3	3.22322911692828	3.5	4	3	4	6	3	4	4	4

Figure 4. Correct values are added to the dataset and an option is provided to download the new dataset

Limitation and future steps

The current version of the toolbox provides an easy way of analyzing data. The toolbox is under active development and can be considered as a useful tool for researchers who wish to understand the data and clean it in a simple and efficient way. We have added several algorithms for outlier detection and gap filling but still there are ways in which we can improve this version. In the future, we would be extending the framework to having a much interactive visualization with plotting options like box plot, correlation plots and also including more features for an in-depth statistical analysis of the data.

Acknowledgements

This work has been supported by the iSCAPE (Improving Smart Control of Air Pollution in Europe) project, which is funded by the European Community's H2020 Programme (H2020-SC5-04-2015) under the Grant Agreement No. 689954.

References

Chen, L.J., Ho, Y.H., Hsieh, H.H., Huang, S.T., Lee, H.C., Mahajan, S., 2018. ADF: An Anomaly Detection Framework for Large-Scale PM2.5 Sensing Systems. IEEE Internet Things J. 5, 559–570.

Eugene Dubossarsky, Tyshetskiy, Y., 2015. Sparse Autoencoder for Automatic Learning of

Representative Features from Unlabeled Data. R Doc.

Ishimaru, S., Weppner, J., Poxrucker, A., Kunze, K., Lukowicz, P., Kise, K., 2014. Shiny.

López-de-Lacalle, J., 2016. tsoutliers R package for detection of outliers in time series. R Doc.

Mahajan, S., Chen, L.J. and Tsai, T.C., 2018. Short-Term PM2. 5 Forecasting Using Exponential Smoothing Method: A Comparative Analysis. Sensors, 18(10), p.3223.

Mahajan, S., Chen, L.J. and Tsai, T.C., 2017, August. An empirical study of PM2. 5 forecasting using neural network. In 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI) (pp. 1-7). IEEE.

Moritz, S., 2018. Time Series Missing Value Imputation. R Doc.

Ottosen, T.-B., Kumar, P., 2019. Outlier detection and gap filling methodologies for low-cost air quality measurements. Environ. Sci. Process. Impacts. 21, 701-713

Ramaswamy, S., Rastogi, R. and Shim, K., 2000, May. Efficient algorithms for mining outliers from large data sets. In ACM Sigmod Record (Vol. 29, No. 2, pp. 427-438). ACM.

Welch, G., Bishop, G., 2006. An introduction to the Kalman filter, Department of University of North Carolina at Chapel Hill: technical report.

Citation for this document:
Mahajan, S., Kumar, P., 2019. Sense Your Data: Sensor Toolbox Manual, Version 1.0. pp.1-7. DOI: 10.13140/RG.2.2.17249.76640/4