Statistical Standard, Methodology and Application

A probability distribution-based approach to impute missing values in hourly PM10 concentration

Rossita Mohamad Yunus

The Study About

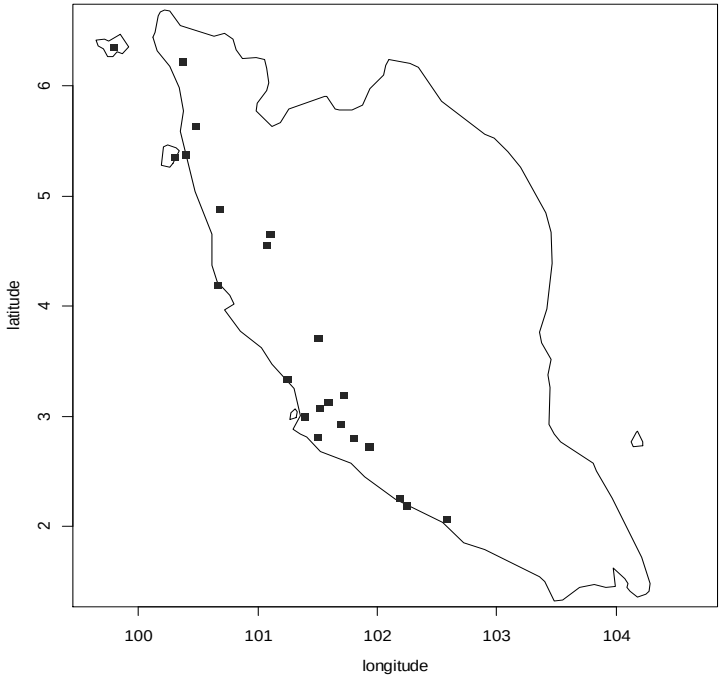
- The study adopt a probability distribution-based approach to impute missing values in hourly PM10 concentration data, and aim to preserve statistical properties similar to that of the observed data.
- The probability based approach is compared to the 4-stations-average (with inverse distance weight) method, which is a well established imputation method that is used in many spatial applications.

Objective of Study

- The main aim is to impute missing values with estimated values that has similar statistical properties to that of the observed data.

The Data

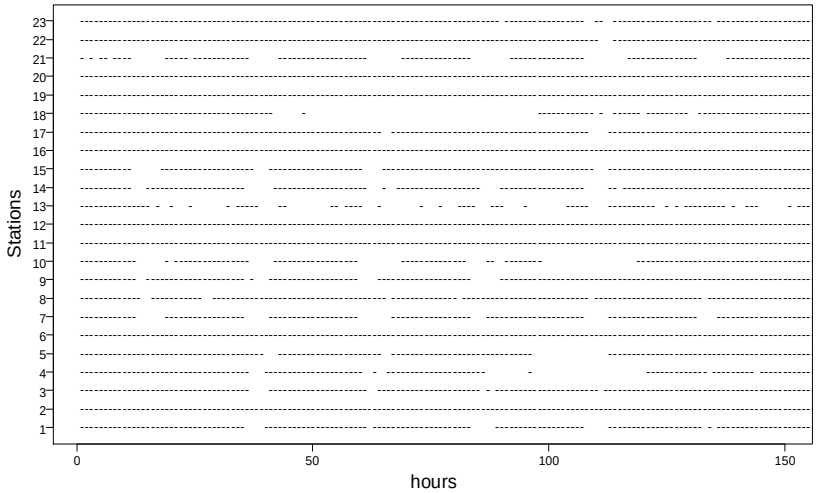
Location of stations



- Hourly PM10 concentration data of twenty three stations from west Peninsular Malaysia
- Data ranges from 1st of January, 2015 to 31st of December 2015 were obtained from the website of Data Terbuka Malaysia

The Gap

Figure shows data coverage for all 23 stations for the first 150 hours.



- none of the stations have data for the entire year, with about half of the studied stations having less than 90 percentage coverage due to gaps in the data

Literature

mean imputation

- missing values are filled by the mean of the known individual data series or known data of the nearest individual data series.
- Intuitively, this method is bias to extreme values that often due to phenomena such as haze episodes.

3-station average

- whereby the gap is filled up with average of data from three nearest stations. (Paulhus and Kohler, 1952),

Literature

k-station weighted average (with inverse distance weighting)

- the where the gaps are estimated using the nearest neighbor stations, with weights are taken as a function of the distances between the studied station and the nearest station (Teegavarapu et al., 2011).
- Teegavarapu and Chandramouli (2005) concluded that, IDW based on four nearest neighbors was the best for the missing data imputation.

Probabilistic approach

- Hasan and Croke (2013) introduced a method for filling gaps in daily rainfall data. In the paper, gaps are filled by data generated randomly from a Poisson gamma distribution, with spatial interpolation for matching the data points.

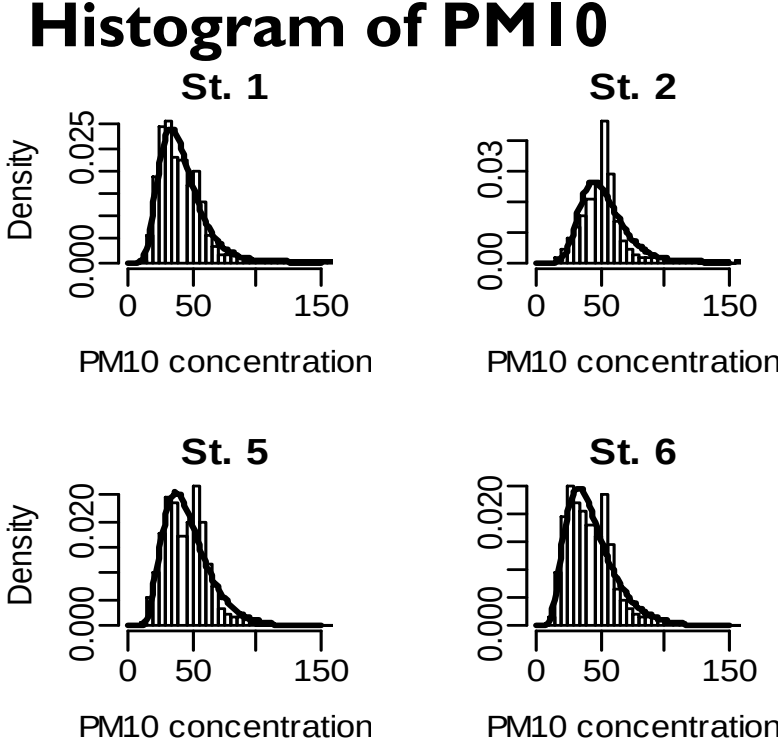
The proposed algorithm

- Let station 'A' has N observations where M missing. The procedures to fill the missing values are performed by the following steps.
- Step 1: Estimate parameters of log normal distribution by fitting the distribution to the available data of the individual station.
- Step 2: Generate M random numbers from the log normal distribution using the parameters obtained in Step 1.

The proposed algorithm (cont.)

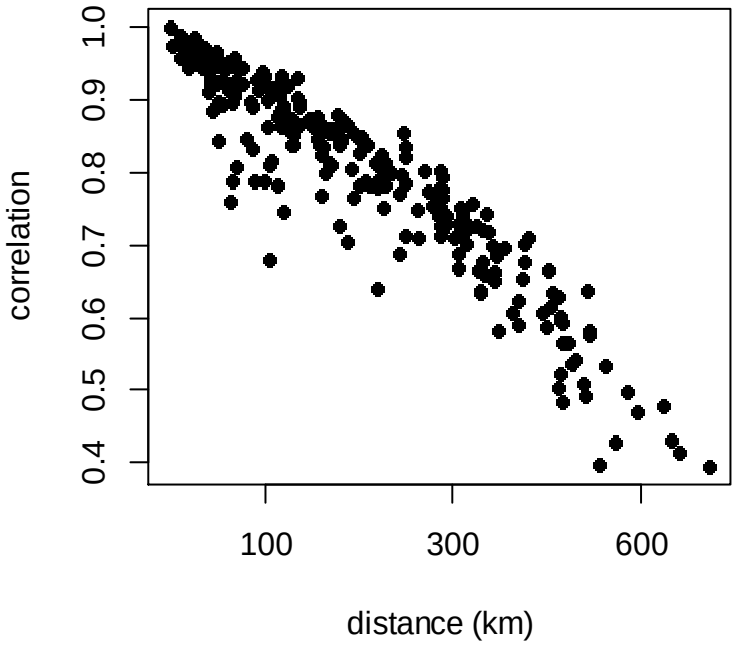
- Step 3: For each missing observations, find weighted average PM10 concentration of four nearest stations. Weights are taken as the inverse of the distances of the neighboring stations from the target station.
- Step 4: If data for all four nearest stations are missing, use the average PM10 concentration of previous hour.
- Step 5: Sort the M generated numbers in Step 2 based on the rank of the average PM10 concentration in Step 3.
- Step 6: Finally, fill the gap with the sorted generated number.

Distribution of Data



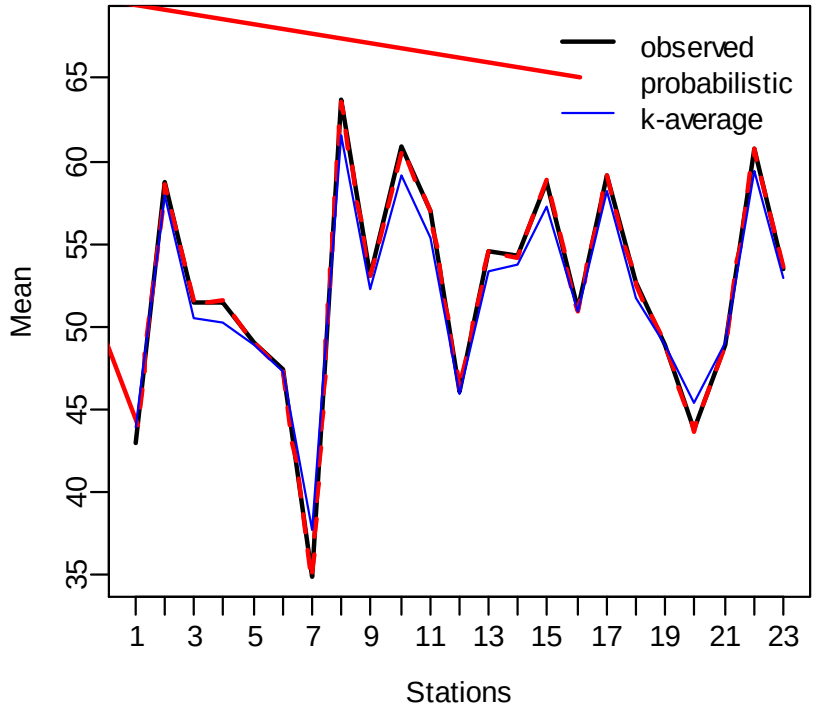
- Figures show Histogram of PM10 concentration for selected stations. Curve line represents the theoretical density function from estimated log normal distribution parameters.
- The log normal distribution fit reasonably well to hourly PM10 concentration data for most studied stations.

Correlation vs. distance



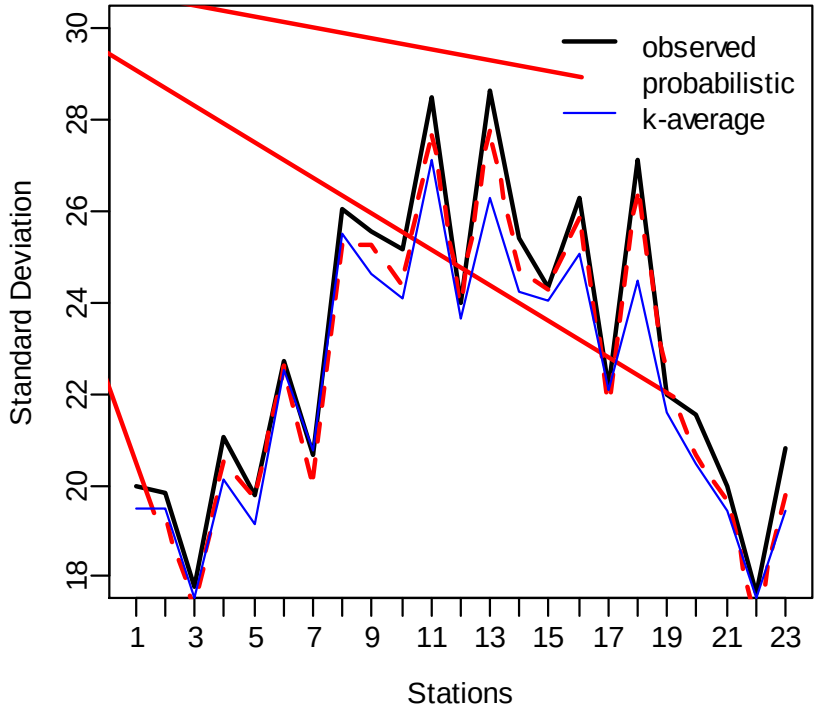
- Figure indicates fairly negative relationship between correlation of PM10 concentration and distances among the stations.
- This suggests that, the imputed value also depends on PM10 concentration data of the nearest stations.

Results 1: Average



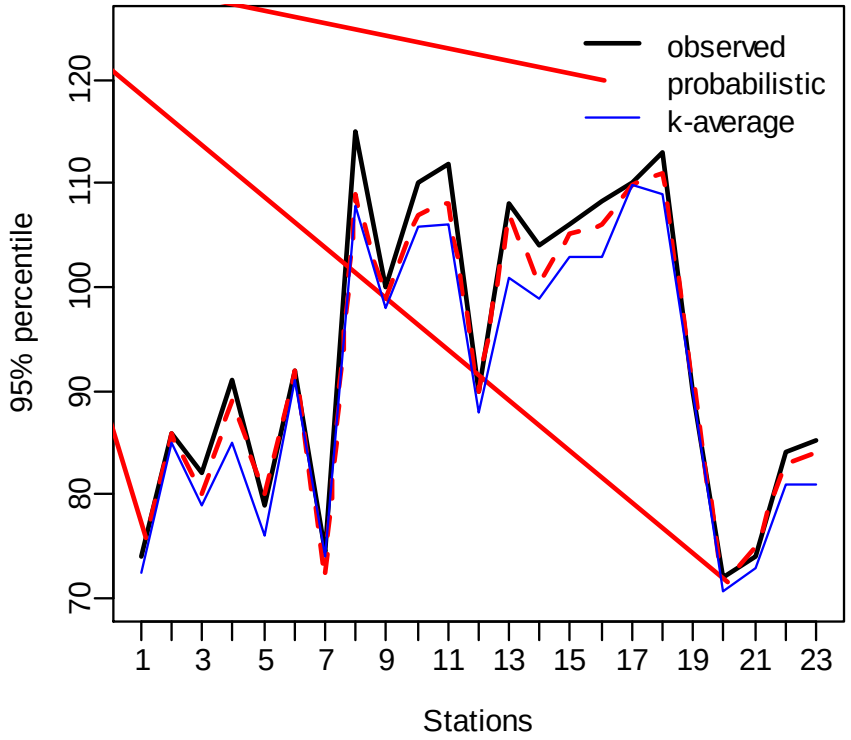
- The average for the probabilistic imputed data is almost similar to that of the observed data.
- The average for the k-station-average imputed data differ slightly to those of the observed and the probabilistic imputed data.

Result 2: Standard deviation



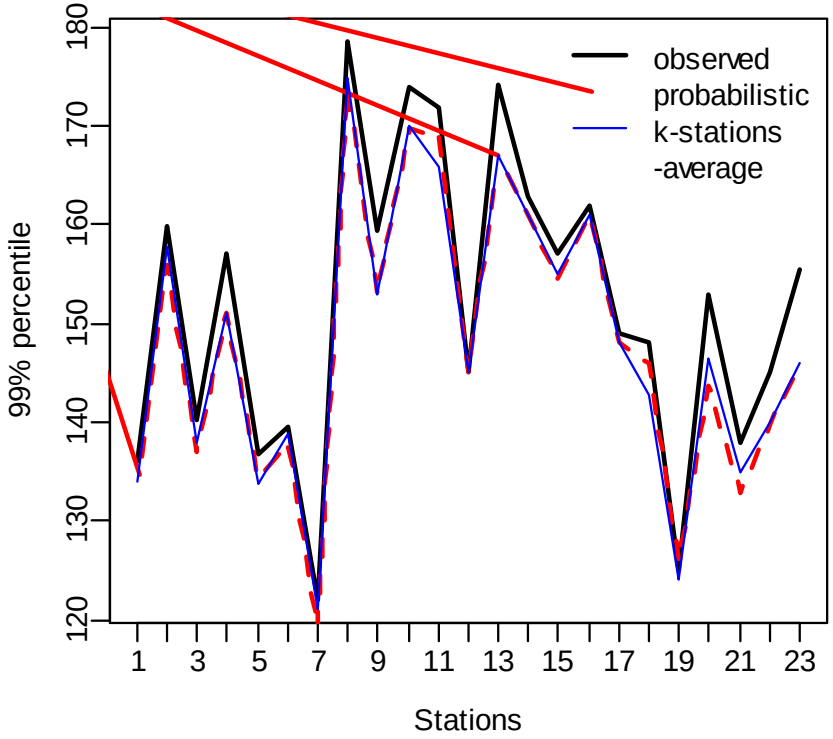
- The standard deviation (variability) for the probabilistic imputed data is closer to that of the observed data, compared to the k-stations-average imputed data.

Result 3: the 95 th percentile



- the probabilistic approach provides better estimation of the 95th percentile than the k-stations-average imputation method, in terms of closer estimation to those of the observed data.

Result 4: the 99th percentile



- Both the probabilistic imputed data and the k-stations-average imputed data underestimates the 99th percentile of the data.

Conclusion

- The paper provides a probability distribution-based approach for imputing the missing values, that considers correlation of PM10 concentration between nearest stations.
- The main aim is to impute with estimated values that has similar statistical properties to that of the observed data.

Conclusion (cont.)

- The probabilistic approach generates data with very similar properties to the observed dataset with respect to the mean, and variability of PM10 concentration. However, the method slightly underestimates the 95th and 99th percentiles in the PM10 concentration data.

Conclusion (cont.)

- The distributional-based imputed data show more similar statistical properties (i.e. mean, standard deviation, and 95th percentile) to the observed data compared to the data imputed using the 4-stations-average (with inverse distance weight) method, which is a well established imputation method that is used in many spatial applications.

The End