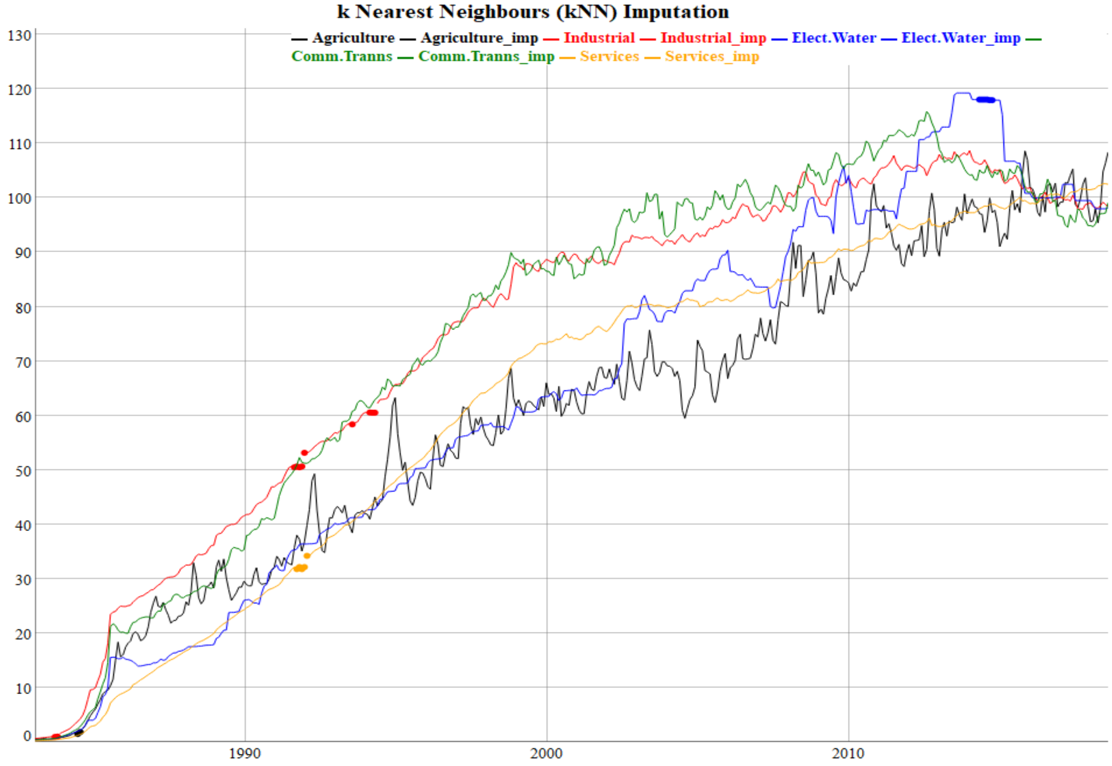
Help on k Nearest Neighbors (kNN) method

This method implements an algorithm that is useful for matching a point with its closest k neighbors in a multi-dimensional space. The assumption behind using kNN for missing values is that a point value can be approximated by the values of the points that are closest to it, based on other variables. It is useful for dealing with all kind of missing data identifies the k-nearest neighbors using Euclidean distance.[1]
Selecting a low k, on one hand will increase the influence of noise and the results will be less generalizable. On the other hand, choosing a high k will tend to blur local effects which are exactly what we are looking for. We use k=5 and the median of aggregated Euclidean distance neighbors of the imputed series, as default parameters. In all imputation figures, the non-NA observations are the colored lines while the imputed values are the colored dots.



References

Kowarik, A., & Templ, M. (2016). Imputation with the R Package VIM. Journal of Statistical Software, 74(7), 1-16.

[1] For any two series x and y with coordinates $(t_1, ...t_N, x_1..., x_N)$ and $(y_1..., y_N)$ where $(t_1...t_N)$ represent the time from 1 to N (on the x axis) and $(x_1...x_N)$ and $(y_1..., y_N)$ are the series values, respectively (on the y axis), the Euclidean distance between x and y at time t_3, for example is: $d(x_3, y_3) = \sqrt{(t_3 - t_3)^2 + (x_3 - y_3)^2} = |x_3 - y_3|$.