We believe the main reason for poor forecast accuracy (with good privacy) in data sets without large groups of similar time series is that the features (and thus, the forecasts) of these series are largely different, and thus identifiable.

I wanted a way to test this hypothesis that was consistent across data sets. Kang, Hyndman, and Li (2020) discuss how linear dimension reduction techniques:

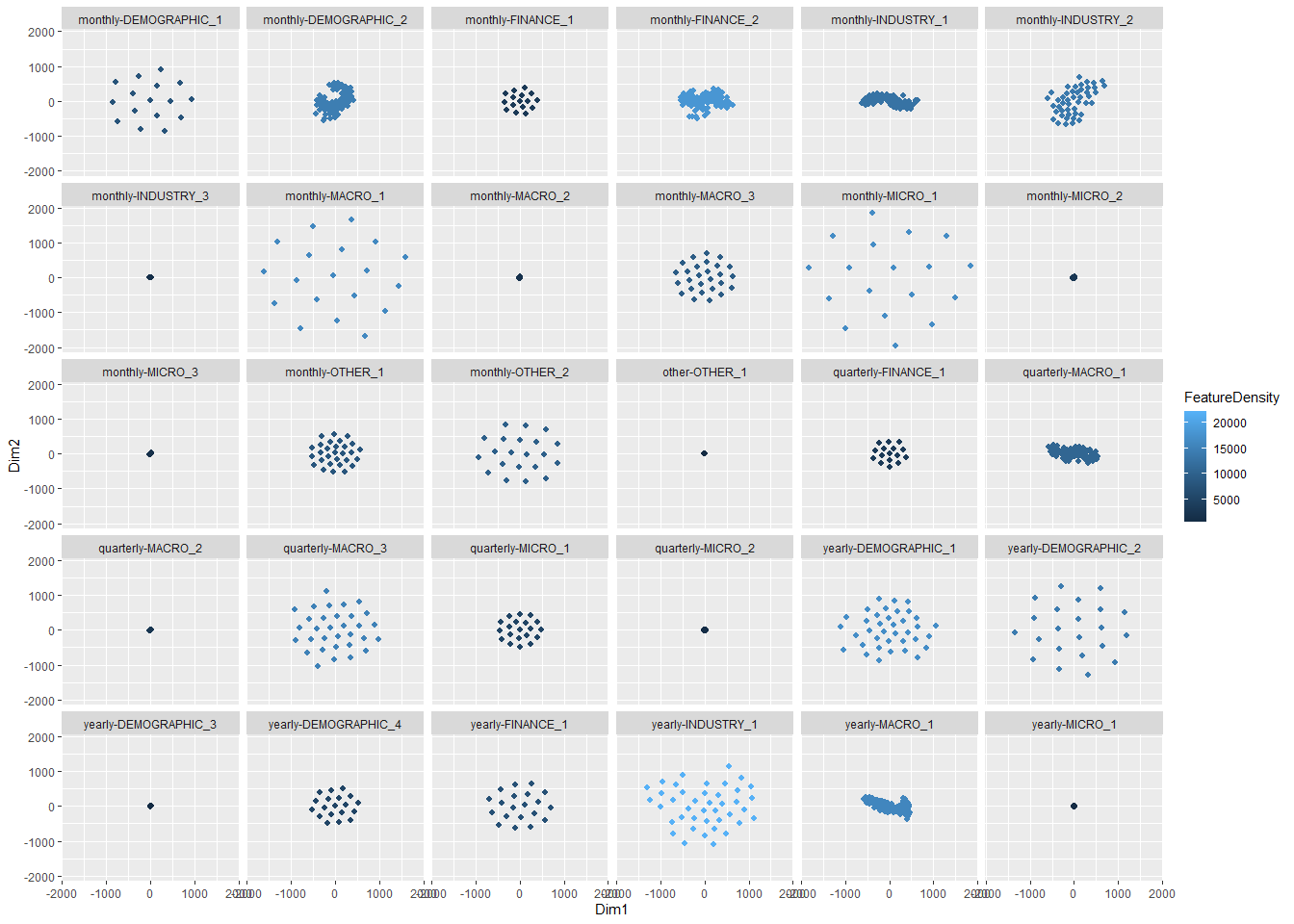
“put more emphasis on keeping dissimilar data points far apart in the low dimensional space. But in order to represent high dimensional data in a low dimensional, nonlinear manifold, it is also important that similar data points are placed close together. **When there are many features and nonlinear correlations are present, using a linear transformation of the features may be misleading**.” (p. 13).

Following this statement, I decided to visualize the time series in each data set using t-SNE. I used the default t-SNE function from the t-SNE package in R. I performed the dimension reduction on the non-whitened feature vectors (since we don’t do any scaling prior to swapping) and I only used the features that RFE selected as most important for each data set. This gave a two-dimensional representation of each feature vector that could then be plotted on the same axes. Visually, we see significantly more variance in the similarity of the feature vectors across data sub(sets) (note that there may be multiple subsets of series for a given data set since we only swap between series that have the same length).

My goal was to come up with a metric for predicting whether our protection method would achieve an acceptable trade-off between privacy and utility. I calculated which is the sum of all pair-wise distances between the dimension-reduced feature vectors, divided by the number of series. In other words, letting denote the set of dimension-reduced feature vectors for data subset ,

,

which can be interpreted as the average pairwise distance within a set of dimension-reduced feature vectors. Higher values of indicate that the series in subset tend to be more dissimilar on the features which were selected as important for maintaining the forecast accuracy of those series. I make this calculation on the t-SNE feature space since this gives a consistent number of dimensions for all data subsets making the are comparable. These distances would not be comparable using the original feature vectors since different numbers of features are selected for swapping in each data set.



**Table 1:** Feature density, average percent change in forecast error, and average percentage of identified series for each data subset.

Kang, Y., Hyndman, R. J., & Li, F. (2020). GRATIS: GeneRAting TIme Series with diverse and controllable characteristics. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *13*(4), 354-376.