A SCALABILITY OVER OTHER DATASETS

A.1 Varying Time Series Size n

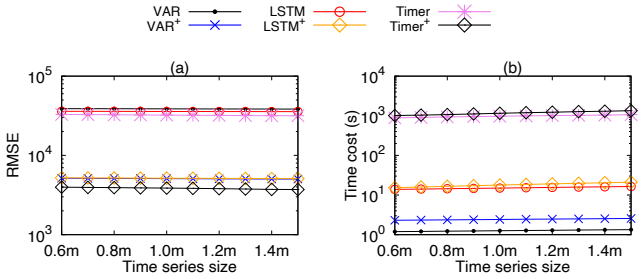

Figure 15: Scalability on time series size n of different regression models over Weather

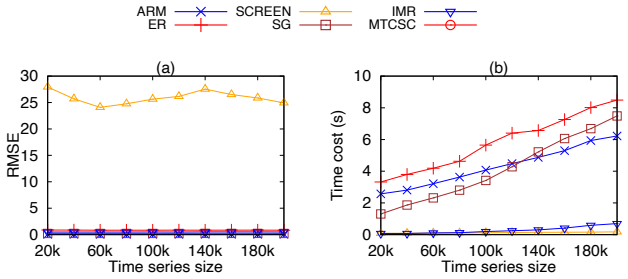

Figure 16: Scalability on time series size n of different cleaning methods over Weather

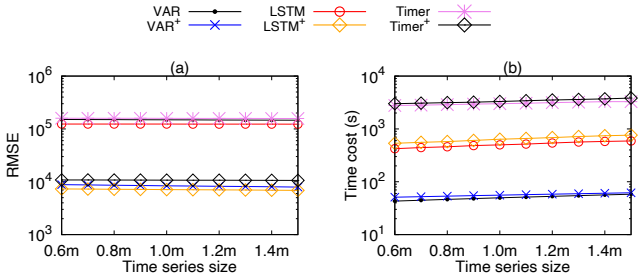

Figure 17: Scalability on time series size n of different regression models over Engine

A.2 Varying Master Data Size m

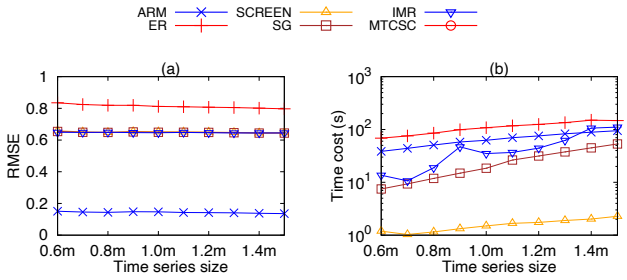

Figure 18: Scalability on time series size n of different cleaning methods over Engine

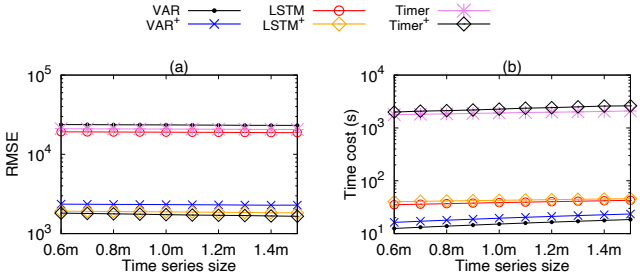

Figure 19: Scalability on time series size n of different regression models over GPS

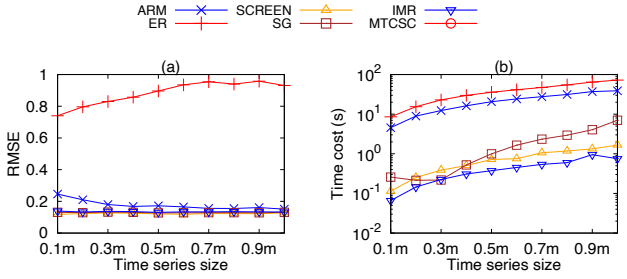

Figure 20: Scalability on time series size n of different cleaning methods over GPS

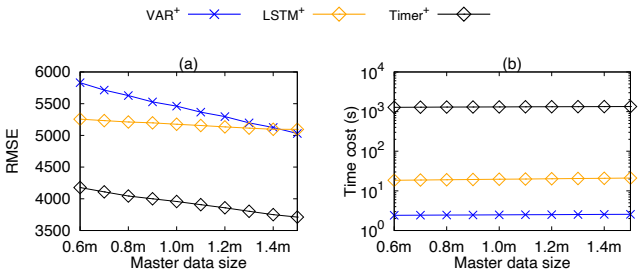

Figure 21: Scalability on master data size m of different regression models over Weather

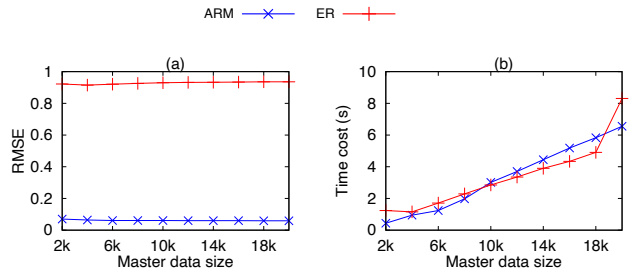

Figure 22: Scalability on master data size m of different cleaning methods over Weather

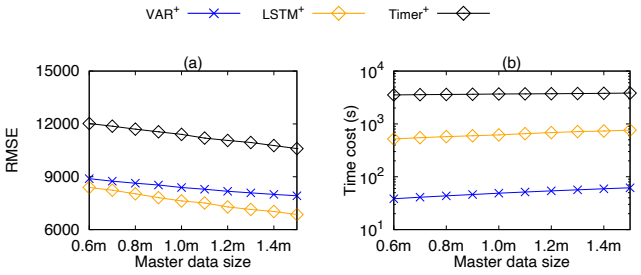

Figure 23: Scalability on master data size m of different regression models over Engine

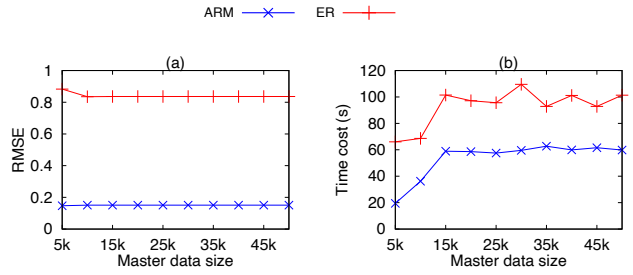

Figure 24: Scalability on master data size m of different cleaning methods over Engine

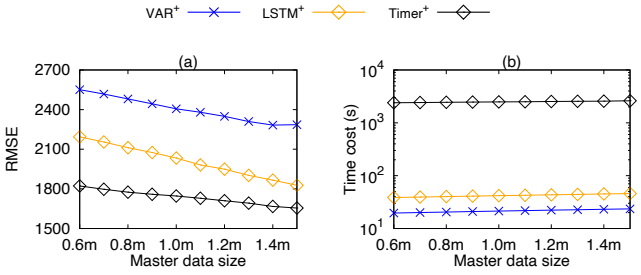

Figure 25: Scalability on master data size m of different regression models over GPS

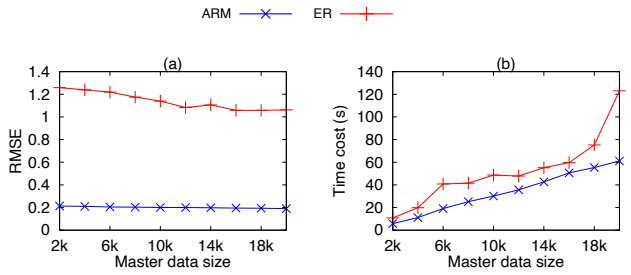

Figure 26: Scalability on master data size m of different cleaning methods over GPS

B ADDITIONAL EXPERIMENTS

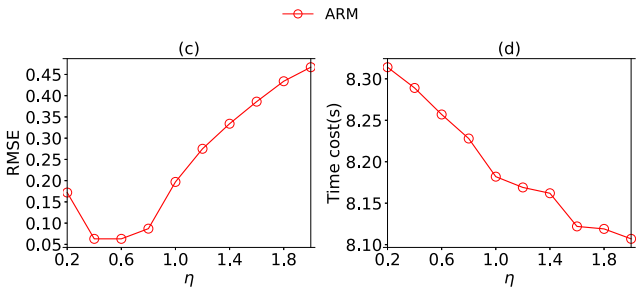

Figure 27: Selecting threshold η of distance to master data over Road

B.1 Varying Master Data Threshold η

The distance threshold η is used to identify errors in our method. The smaller the distance threshold is, the more likely it is for a data point to be identified as an error. That is, the distance between the point and its nearest neighbor in master data could exceed the threshold. If η is too small, e.g., 0.2 in Figure 27, too many points will be regarded as errors, and thus excessively repaired. The error repairing RMSE and anomaly detection precision are thus low. On the other hand, a too large η may fail to identify errors, leading to low repair RMSE and anomaly detection performance as well. The time cost generally decreases with the increase of η, since less points will be processed as errors. Note that the existing method ER does not have such a parameter on distance to master data and thus is omitted.

Intuitively, to determine a proper η, we observe the distribution of distances between data points and their nearest neighbors in the master data. Threshold η is determined by the k-sigma rule. For instance, Figure 27 suggests a threshold $\eta = \mu + \sigma = 0.7$, referring to one-sigma in the Road dataset.