# Workflow for Data Preprocessing

Accuracy >= threshold (e.g. 80%)

Google-Drive

sms.xml

Convert to CSV

Use in STATA

./SMSBuckupRestore

Daily backups

sms\_”date”.csv

Pre-process

Fix timestamps, remove duplicates etc.

Regularize (time-interval = 1 hr.)

sms\_observed.csv

Identify missing events (type of event = message, method = ‘strict’, ‘relaxed’, etc.)

sms\_rect\_hr.csv

Regularized dataset with label for missing

OneDrive:

../processed\_sms/

OneDrive:

../raw\_sms/

Evaluate imputation model (model = ‘moving average, ‘random forest’ etc)

sms\_rect\_hr.csv

NO

Impute data (model = ‘model’, input\_data = sms\_”date”\_v2.csv)

Yes

sms\_rect\_hr\_imputed.csv

Regularized with missing events imputed

Figure 5: Workflow for data preprocessing