Evaluating Floating Car Data Quality

for Knowledge Discovery

Vítor Cerqueira[1], Luís Moreira-Matias[2], and Jihed Khiari[2]

U.Porto, Portugal[1]

NEC Laboratories Europe,Germany[2]

September, 2017

Outline

Outline

Motivation

Increasing availability of GPS-enabled vehicles and Floating Car Data (FCD) sources – Most fleets will be published as open data

Data quality issues

- GPS devices with different protocols (3G/4G);
- Road/urban topology (e.g. narrow streets, high buildings)

Costs of using unreliable FCD

- Suboptimal traffic control actions;
- Excessive storage usage;
- Excessive computational power usage.

Main goal

Develop an automated analysis of FCD sources

- Efficiently discard unreliable ones

Approach

- Set of statistical indicators
  - parametrizable, yet interpretable and with standard output
- Spatio-temporal coverage, and veracity of FCD

Main goal

Develop an automated analysis of FCD sources

- Efficiently discard unreliable ones

Approach

- Set of statistical indicators
  - parametrizable, yet interpretable and with standard output
- Spatio-temporal coverage, and veracity of FCD

Current state of the art

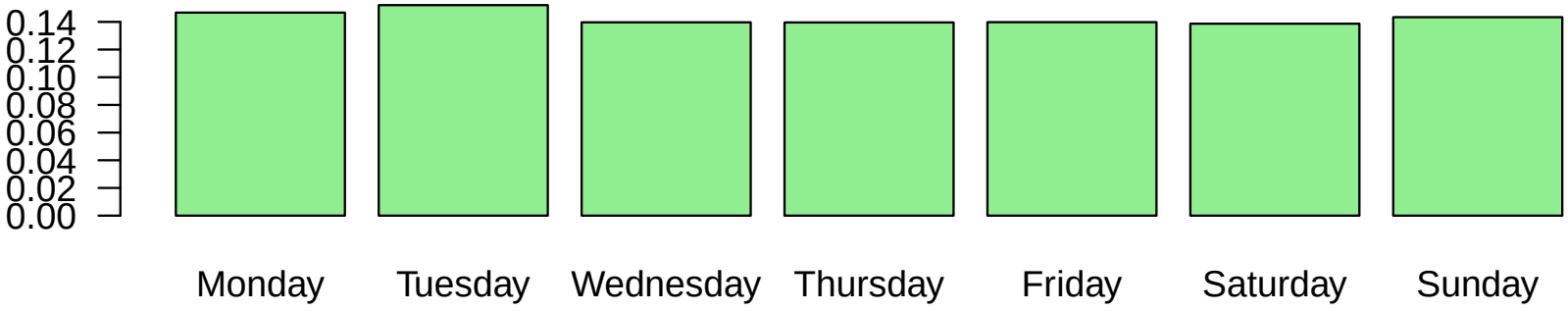
- Mostly about GPS accuracy analysis
- Used as pre-processing step, without a standard and comparable output.

Outline

Granularity - frequency of transmission of GPS traces

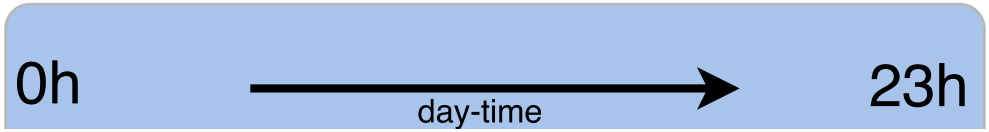
- Relevant for tracking vehicles, and tasks such as map-matching and congestion prediction

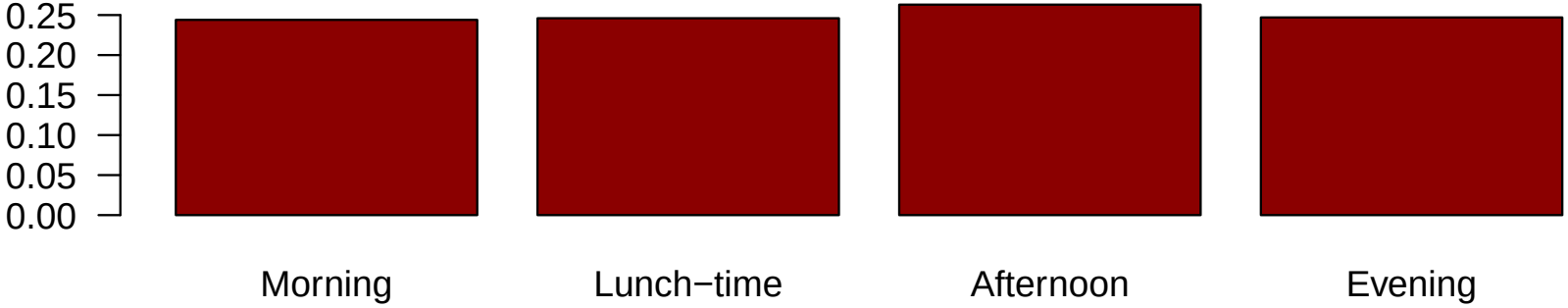
Macro temporal coverage



- Relevant for demand forecasting tasks

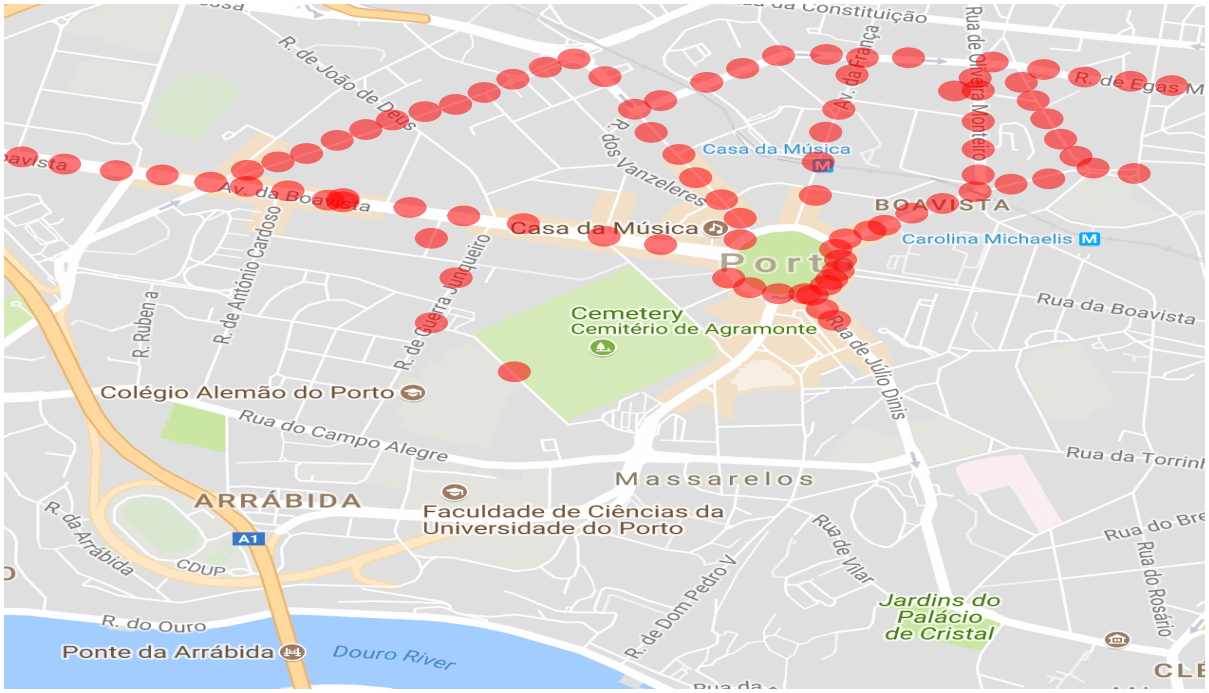
Micro temporal coverage





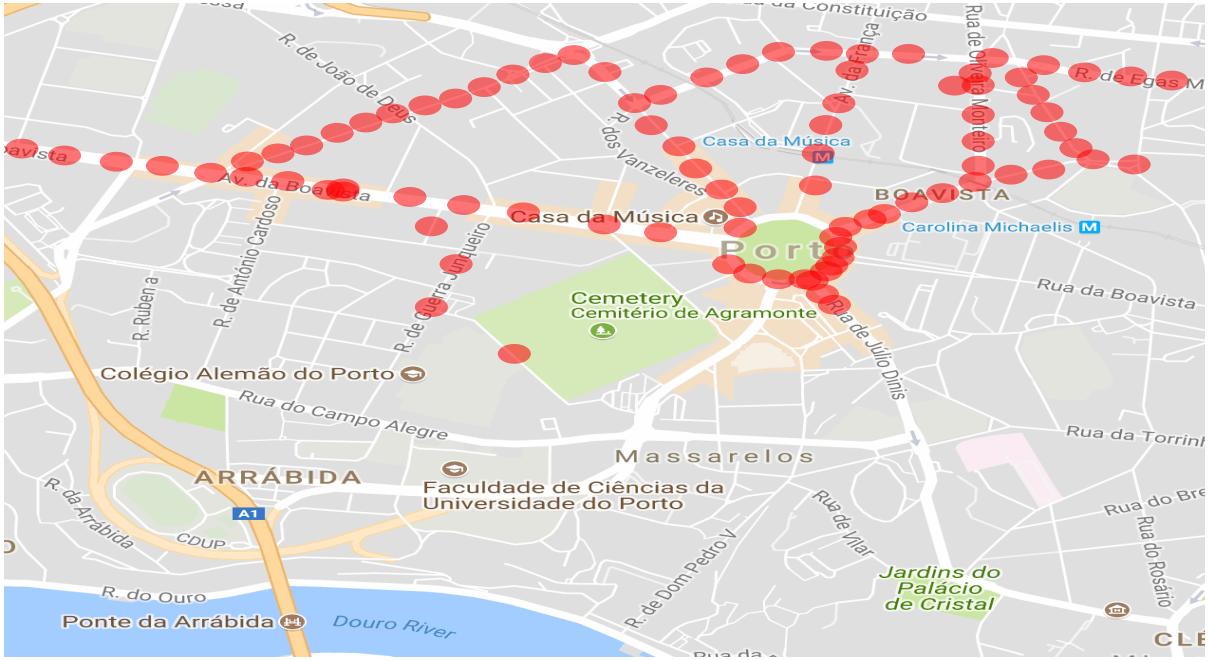
- Also relevant for demand forecasting tasks, specifically in-day seasonality: Rush hours, demand peaks due to special events (e.g. soccer match)

Spatial coverage



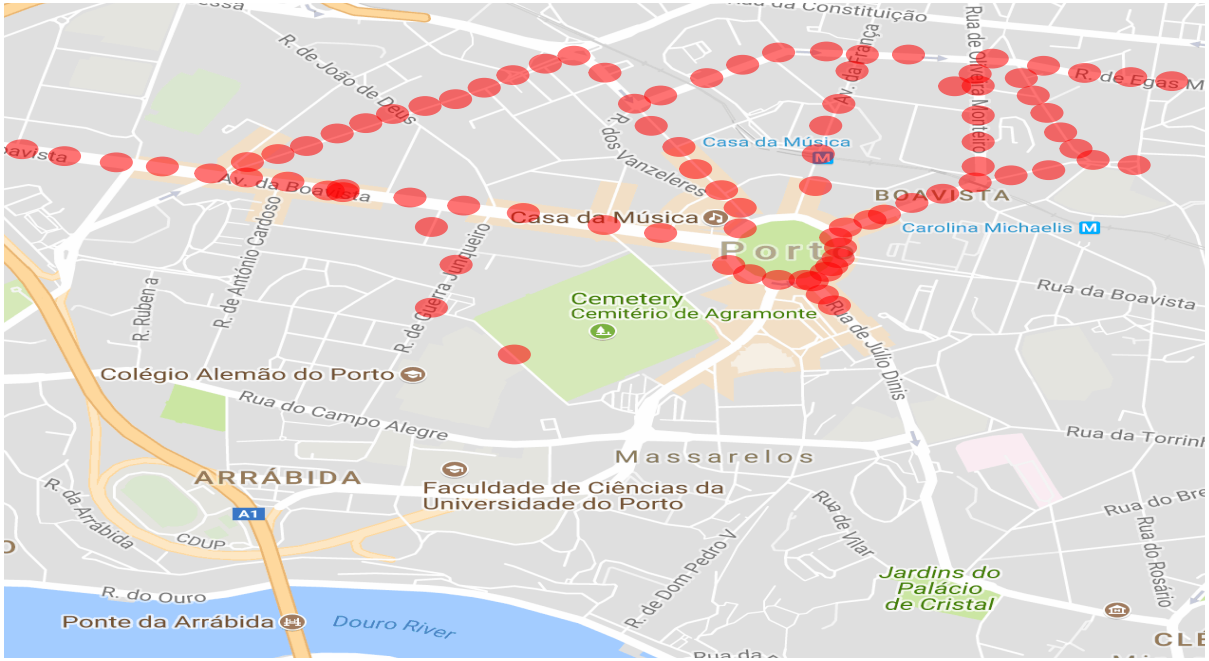
- Some parts of a town a more relevant than others
  - Road density;

Spatial coverage



- Some parts of a town a more relevant than others
  - Road density;
  - Proximity to landmarks;

Spatial coverage



- Some parts of a town a more relevant than others
  - Road density;
  - Proximity to landmarks;
  - Neighborhood

Missing data

- Large time intervals without GPS transmission
  - How many gaps exist per trip, on average;
  - How are they typically spread across a trip

Reliability

- Analysis of objectivity of the dataset;
  - Reachability;
  - Activity

Accuracy

- Measuring the accuracy of GPS devices

Outline

- **Timestamp**, **Vehicle Id**, **Trip Id**, **Latitude**, and **Longitude**

Table: Nanjing & San Francisco taxi fleets

	Nanjing	**San Francisco**
No. GPS traces	18 million	11 million
No. Trips	432.899	959.025
No. vehicles	7648	536
Timespan	1 day	23 days
Median Trip Duration	10 min	7.5 min

- Data mining tasks:
  - Travel time prediction;
  - Origin-destination flow estimation

Table: Results from the Data Mining Experiment I

Error	Nanjing	San Francisco
RMSE	1619.34	624.96
MAE	700.7	455.2

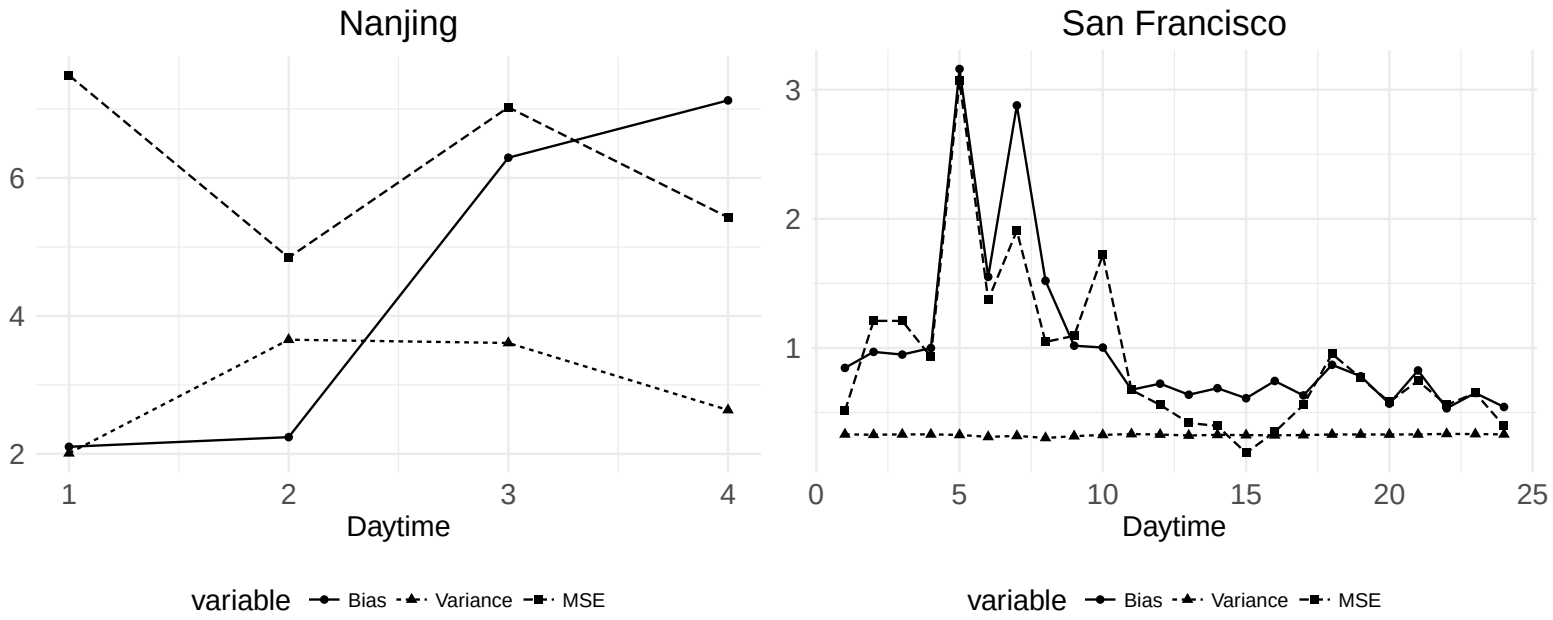


Figure: MSE bias-variance decomposition for DM-II

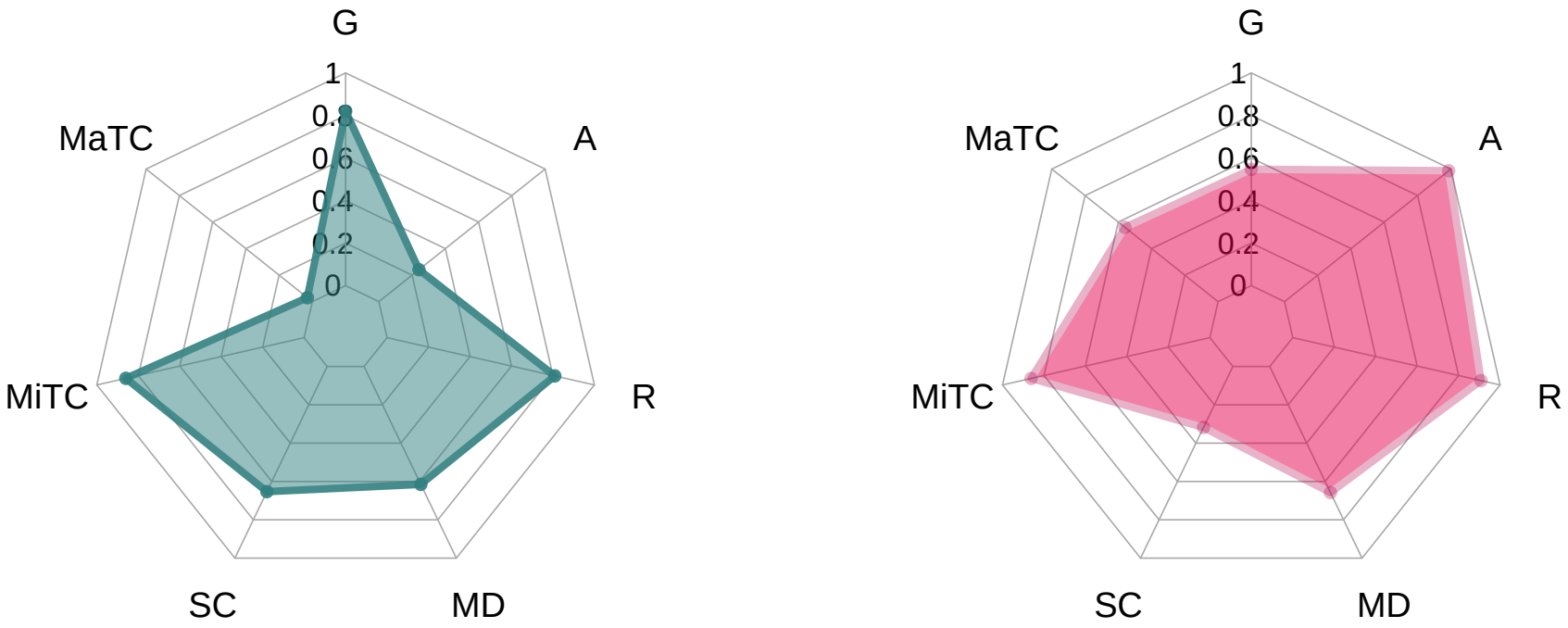


Figure: Methodology results for Nanjing and San Francisco

Results

- **Macro Temporal Coverage** shows the greatest discrepancy:
  - Driven by the timespans of the datasets
  - Suggests possible generalization issues in Nanjing, corroborated by the results of the data mining tasks.

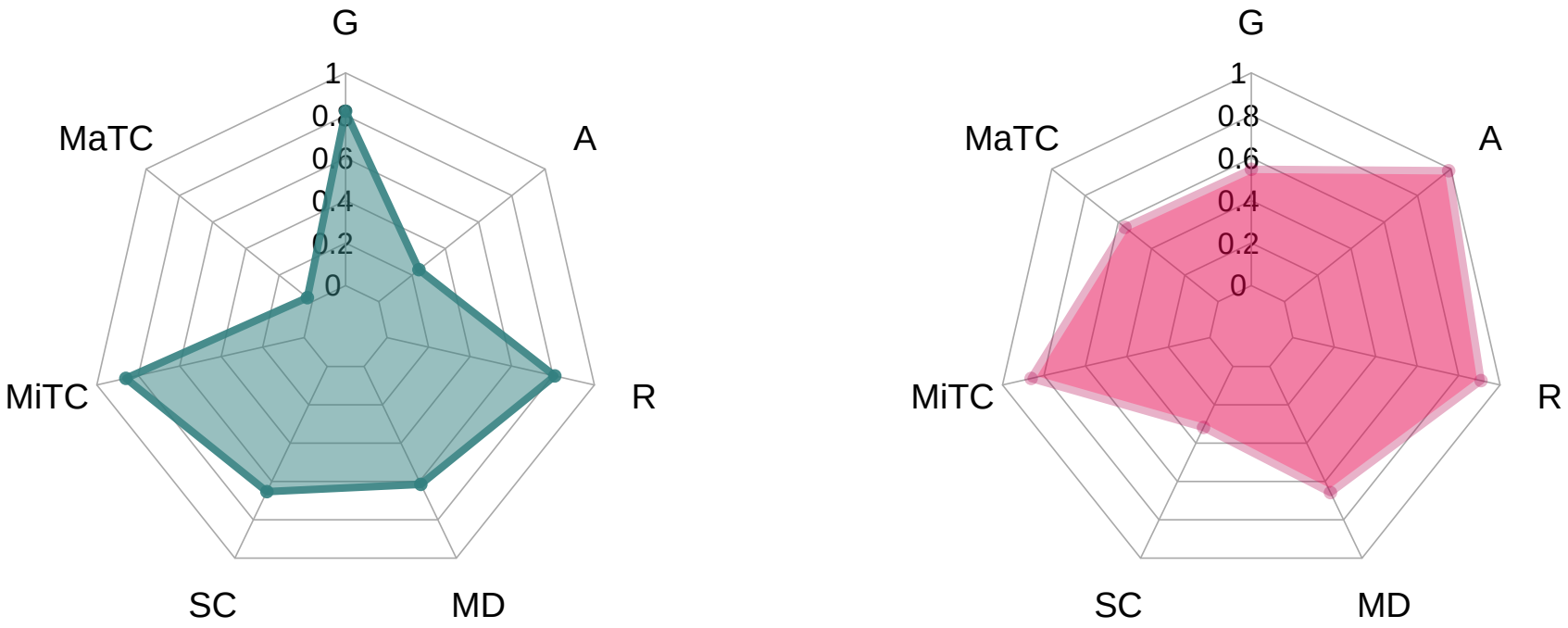


Figure: Methodology results for Nanjing and San Francisco

Results

- **Spatial Coverage** difference is also noticeable:
  - Nanjing's cover more ground in less time, but with a much larger fleet;
  - San Francisco's value suggests that the fleet operates only in a small part of the town.

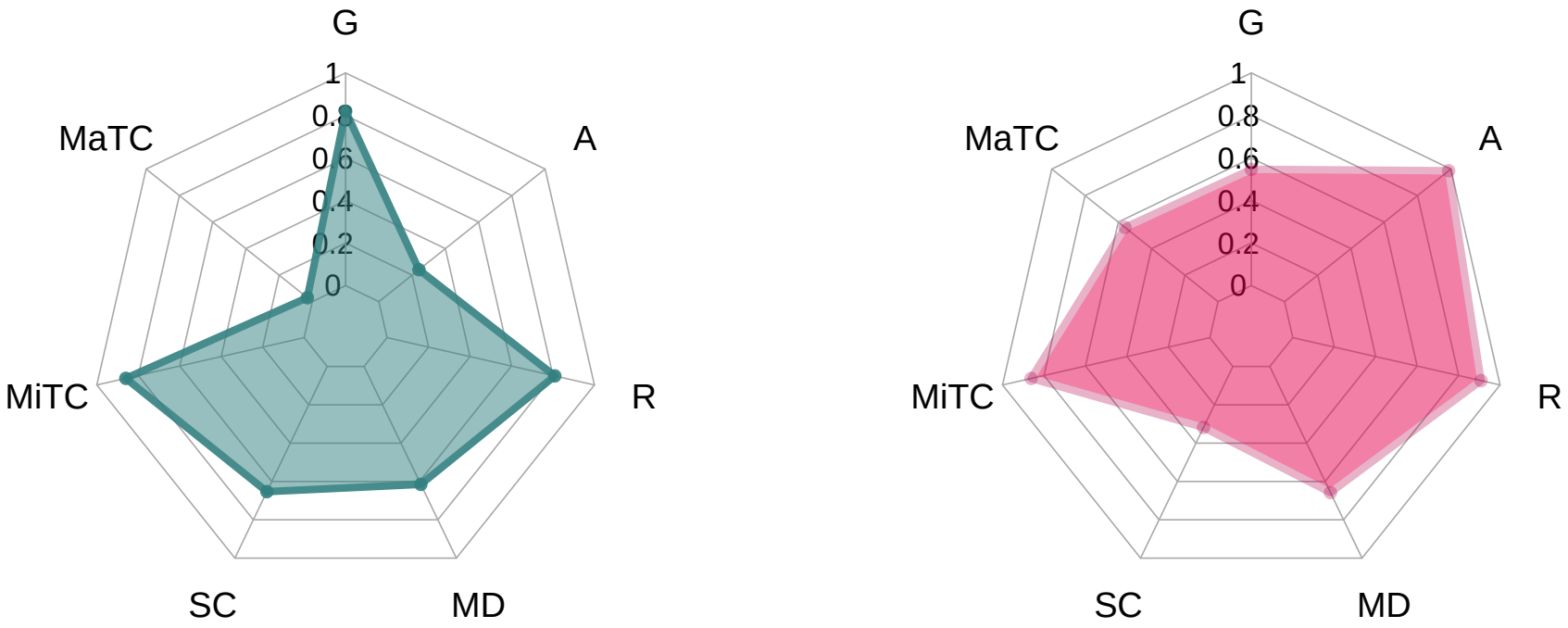
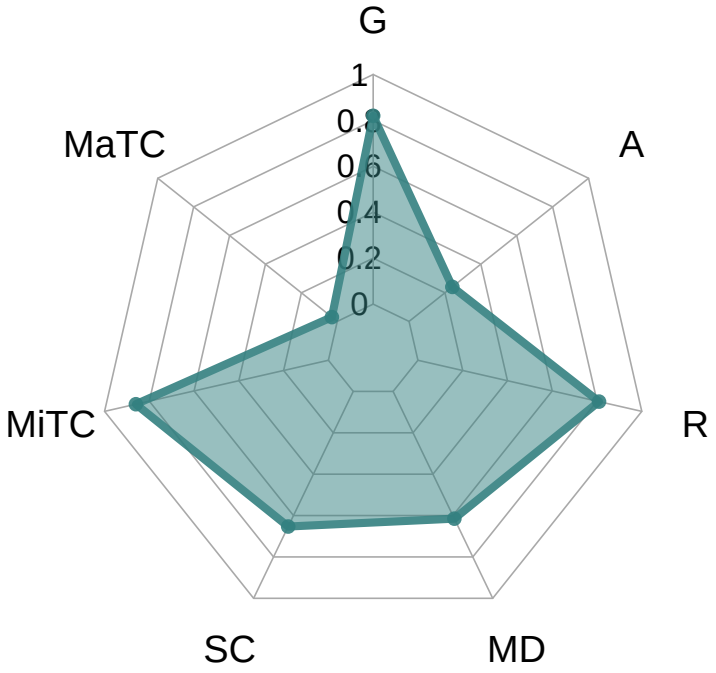


Figure: Methodology results for Nanjing and San Francisco

Results

- **Missing data** is comparable:
  - Nanjing shows a better granularity;
  - But it also has an higher no. of missing observations per trip, on average.



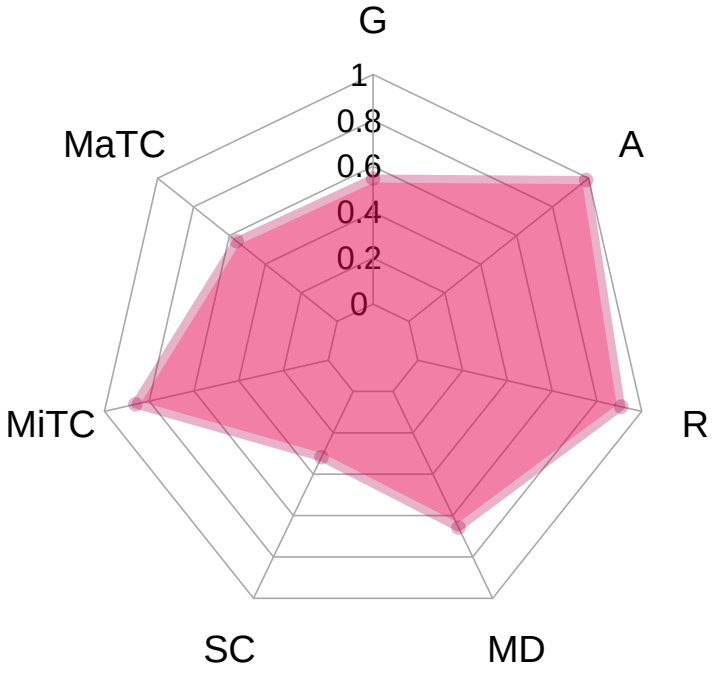


Figure: Methodology results for Nanjing and San Francisco

Results

- **Accuracy** is higher for San Francisco:
  - Average error of 40m and 15m, for Nanjing and San Francisco;

Outline

Results

- Link between the results of the methodology and the results of the data mining tasks
  - a way to reason about the differences in performance of the same model under different scenarios
- an automated approach without conducting any manual data mining procedure

Applicability

- Nanjing ×
- San Francisco ✓

Potential deployment

Market value

- Emerging data market
  - A standardized evaluation framework is helpful to define the market value of FCD feeds

Potential deployment

Market value

- Emerging data market
  - A standardized evaluation framework is helpful to define the market value of FCD feeds

Troubleshooting

- Reliability of FCD
  - Detecting malfunctions on the sensors

Potential deployment

Market value

- Emerging data market
  - A standardized evaluation framework is helpful to define the market value of FCD feeds

Troubleshooting

- Reliability of FCD
  - Detecting malfunctions on the sensors

Saving communicational broadband

- Reducing the storage requirements of typical TMC visualization systems
  - Periodically analysis of the data broadcasted by the input sources
- Reducing computational demand
  - for short-term inference of the traffic status using data mining

Thank you.