Data Driven Methods for Train Delay Prediction

TIONG KAH YONG, LUND UNIVERSITY

Content

Introduction

Research design

Answering research questions

Conclusion and future research

LUND UNIVERSITY

Introduction

- Train delay is defined as the *deviation of actual train events from scheduled train events.*

- Train delay is defined as the *deviation of actual train events from scheduled train events.*

- *High-capacity utilisation* and *heterogeneous traffic* make the railway network susceptible to delays.

LUND UNIVERSITY

Train delay prediction model

- *Predicting* the expected train traffic conditions at a future time

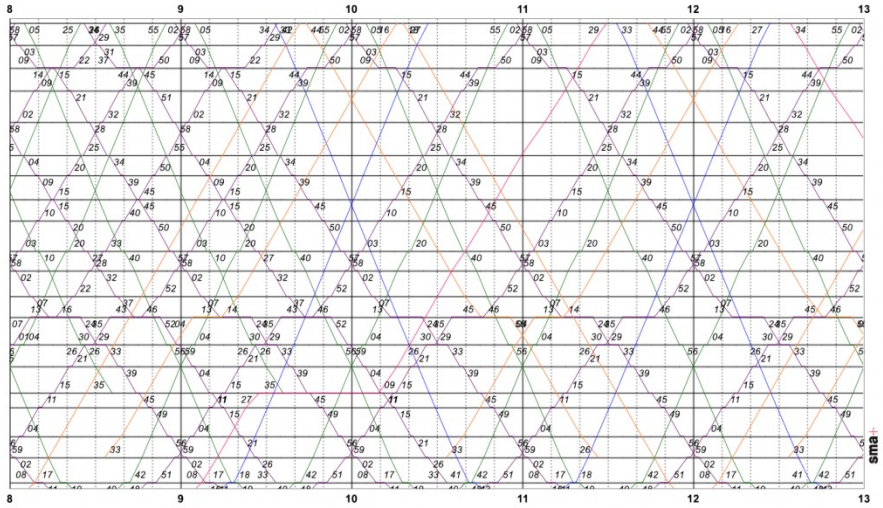
Train delay prediction model

- *Predicting* the expected train traffic conditions at a future time.

- *Input* for solving many problems related to train traffic management.

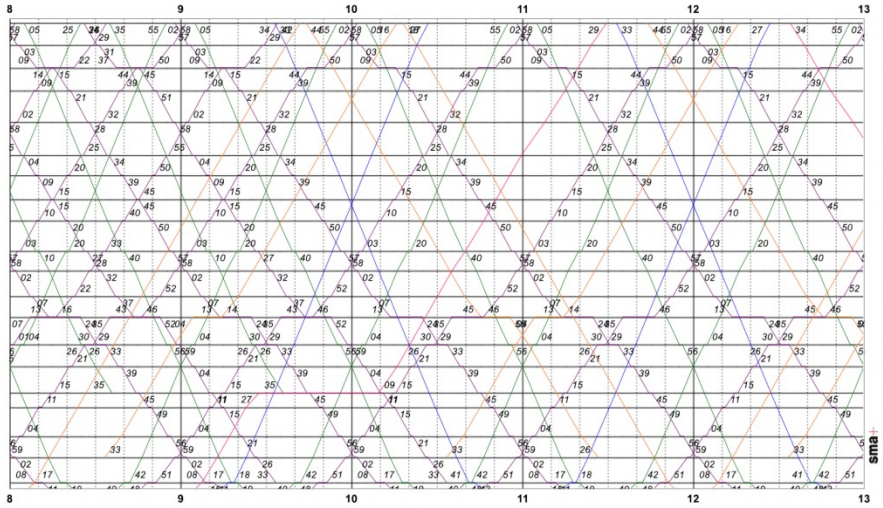
Train delay prediction model

Train delay prediction model



Timetable planning

Train delay prediction model

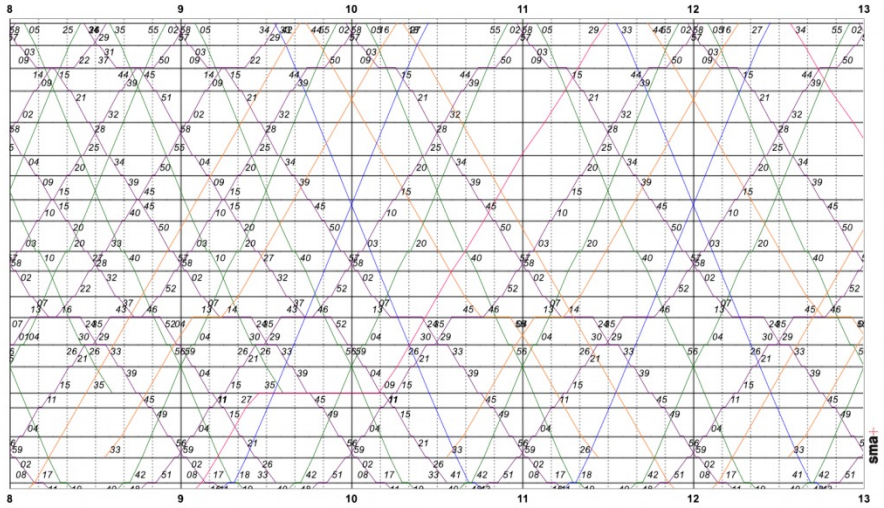


Timetable planning

Real-time train management

LUND UNIVERSITY

Train delay prediction model



Timetable planning

Real-time train management

Reliable passenger information system

LUND UNIVERSITY

Research design

Research gap

Research gap

- *Insufficient understanding* of existing train delay prediction models.

LUND UNIVERSITY

Research gap

- *Insufficient understanding* of existing train delay prediction models.

- There is a lack of innovation in developing train delay prediction models with *practical applications*.

LUND UNIVERSITY

Research aim

To increase understanding of data-driven train delay prediction models

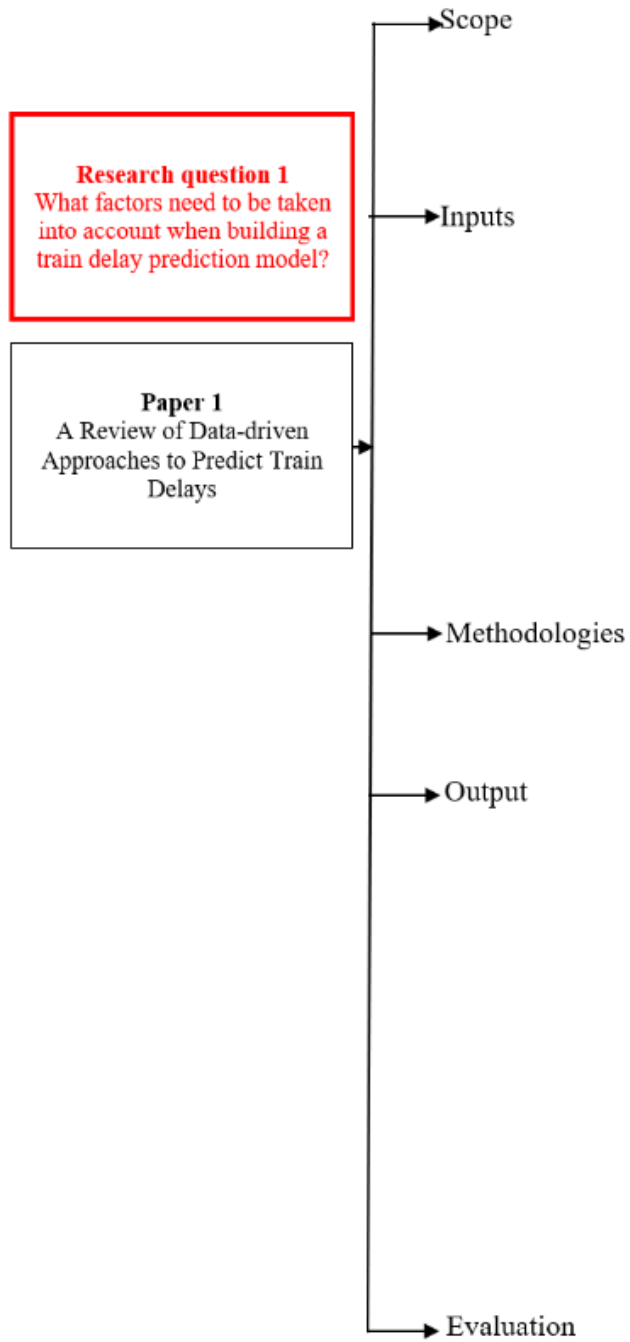
Research questions, papers and their connections

Research question 1
What factors need to be taken into account when building a train delay prediction model?

To increase *our understanding of the various aspects* that must be considered.

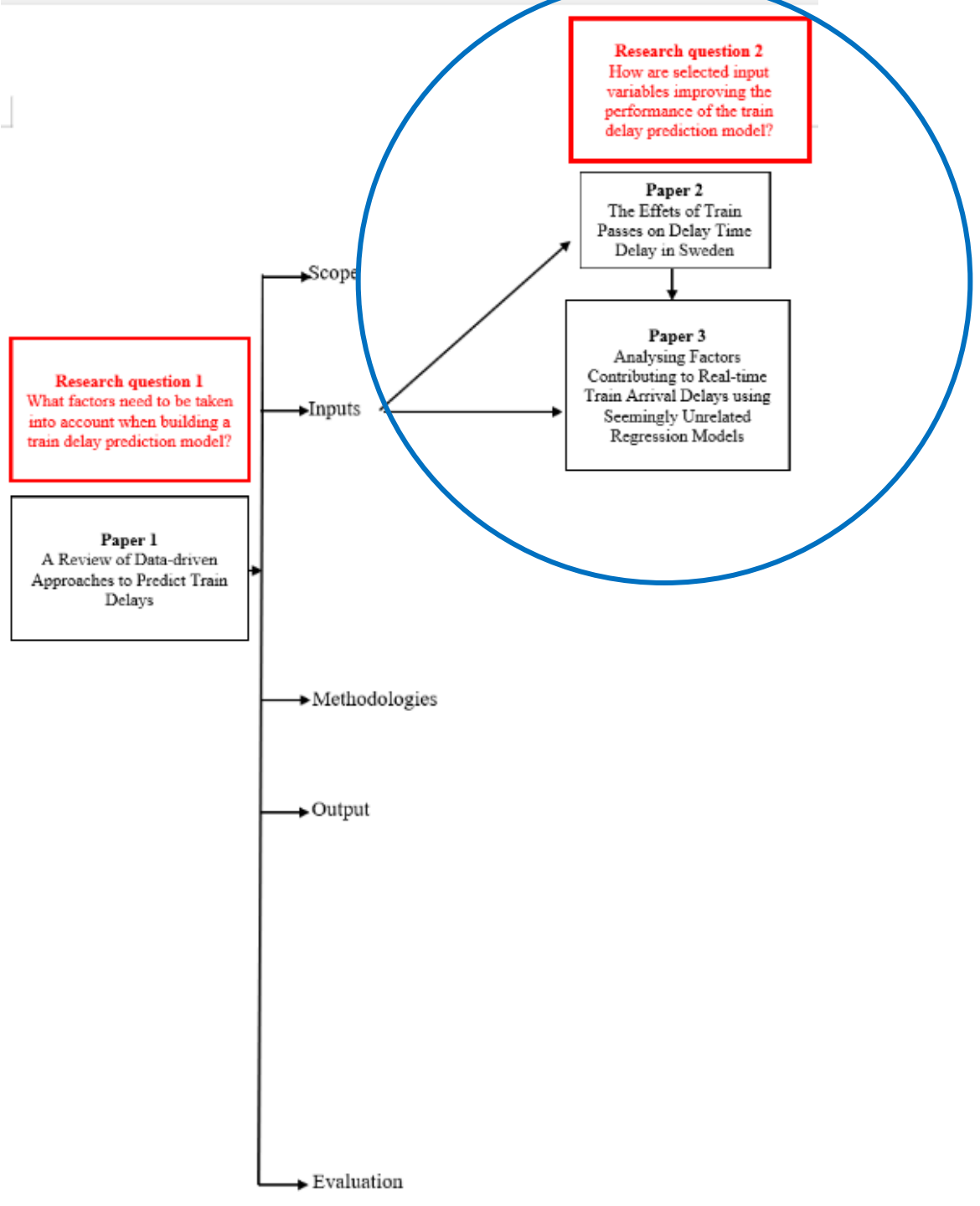
Research question 1
What factors need to be taken into account when building a train delay prediction model?

Paper 1
A Review of Data-driven Approaches to Predict Train Delays


Research question 1
What factors need to be taken into account when building a train delay prediction model?
Paper 1
A Review of Data-driven Approaches to Predict Train Delays
Scope
Inputs
Methodologies
Output
Evaluation

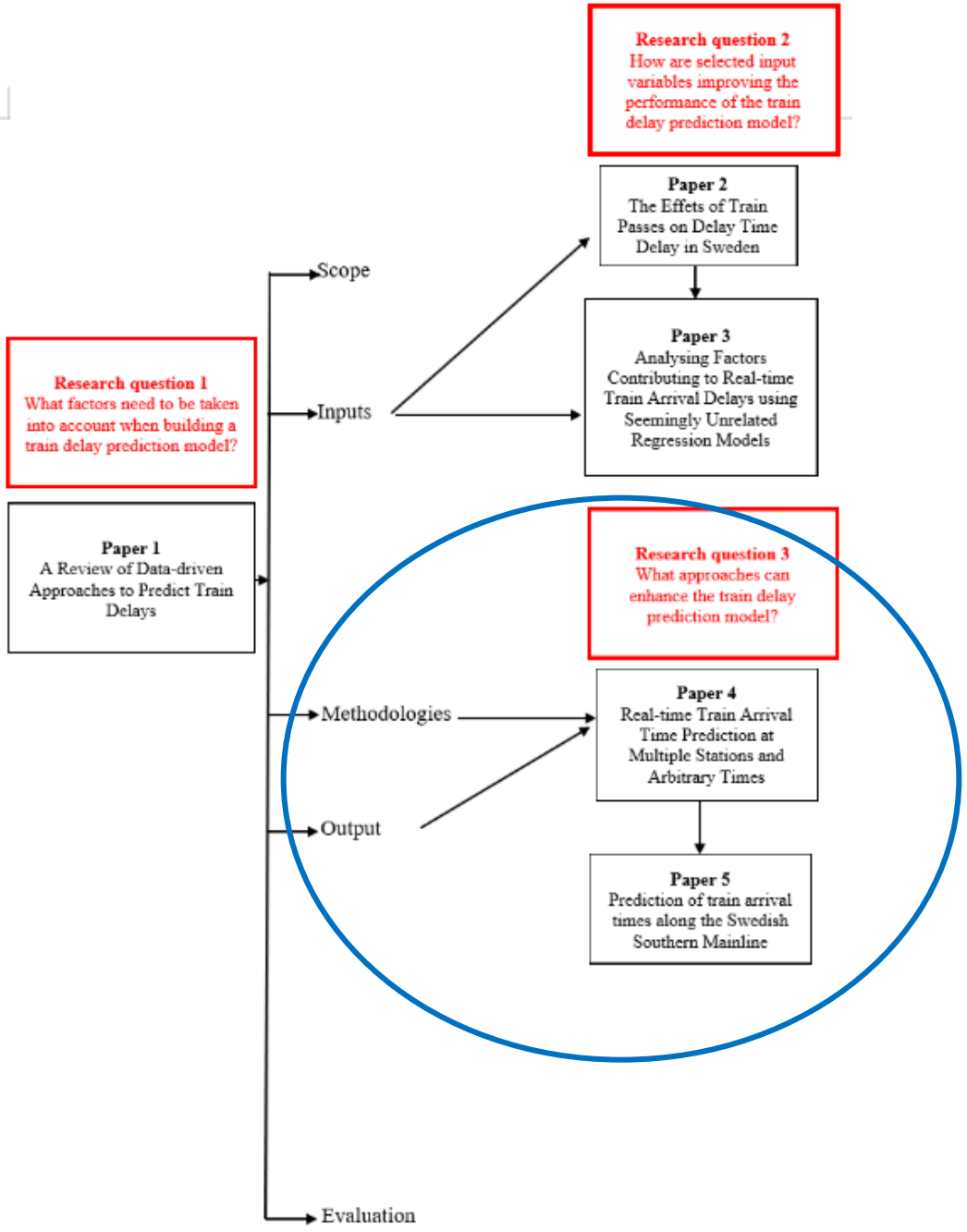
Research question 2
How are selected input variables improving the performance of the train delay prediction model?

To *identify useful input variables* to enhance model performance


Research question 2
How are selected input variables improving the performance of the train delay prediction model?
Paper 2
The Effets of Train Passes on Delay Time Delay in Sweden
Paper 3
Analysing Factors Contributing to Real-time Train Arrival Delays using Seemingly Unrelated Regression Models
Scope
Research question 1
What factors need to be taken into account when building a train delay prediction model?
Inputs
Paper 1
A Review of Data-driven Approaches to Predict Train Delays
Methodologies
Output
Evaluation

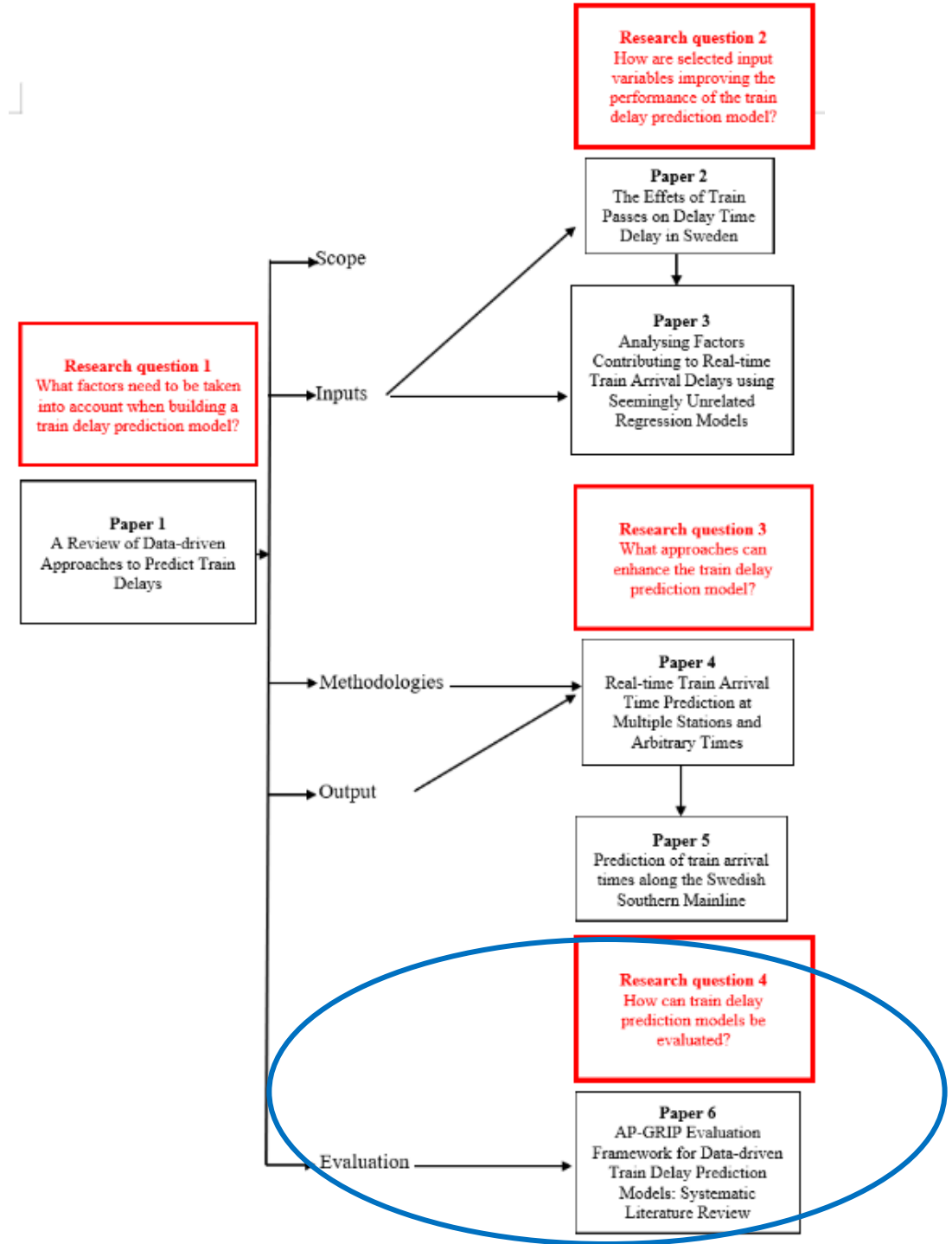
Research question 3
What approaches can enhance the train delay prediction model?

The *formulation of innovative technical solutions* to address the current modelling challenges


Research question 2
How are selected input variables improving the performance of the train delay prediction model?
Paper 2
The Effets of Train Passes on Delay Time Delay in Sweden
Scope
Paper 3
Analysing Factors Contributing to Real-time Train Arrival Delays using Seemingly Unrelated Regression Models
Research question 1
What factors need to be taken into account when building a train delay prediction model?
Inputs
Paper 1
A Review of Data-driven Approaches to Predict Train Delays
Research question 3
What approaches can enhance the train delay prediction model?
Methodologies
Paper 4
Real-time Train Arrival Time Prediction at Multiple Stations and Arbitrary Times
Output
Paper 5
Prediction of train arrival times along the Swedish Southern Mainline
Evaluation

Research question 4
How can train delay prediction models be evaluated?

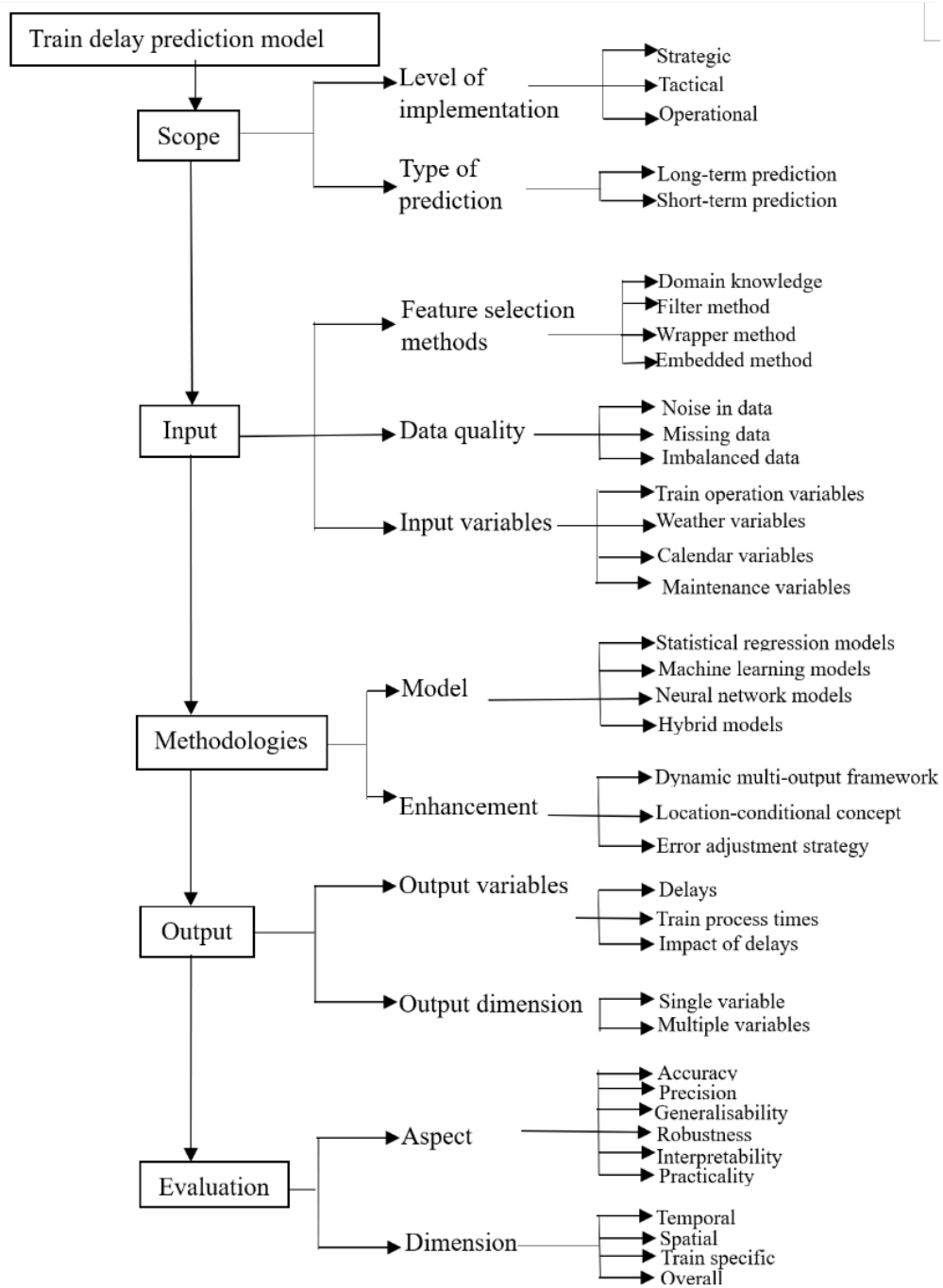
To thorough assessment of train delay prediction models


Research question 2
How are selected input variables improving the performance of the train delay prediction model?
Paper 2
The Effets of Train Passes on Delay Time Delay in Sweden
Scope
Research question 1
What factors need to be taken into account when building a train delay prediction model?
Inputs
Paper 3
Analysing Factors Contributing to Real-time Train Arrival Delays using Seemingly Unrelated Regression Models
Paper 1
A Review of Data-driven Approaches to Predict Train Delays
Research question 3
What approaches can enhance the train delay prediction model?
Methodologies
Paper 4
Real-time Train Arrival Time Prediction at Multiple Stations and Arbitrary Times
Output
Paper 5
Prediction of train arrival times along the Swedish Southern Mainline
Research question 4
How can train delay prediction models be evaluated?
Evaluation
Paper 6
AP-GRIP Evaluation Framework for Data-driven Train Delay Prediction Models: Systematic Literature Review

Answering research questions

Research question 1

What factors need to be taken into account when building a train delay prediction model?


Train delay prediction model
Scope
Level of implementation
Strategic
Tactical
Operational
Type of prediction
Long-term prediction
Short-term prediction
Input
Feature selection methods
Domain knowledge
Filter method
Wrapper method
Embedded method
Data quality
Noise in data
Missing data
Imbalanced data
Input variables
Train operation variables
Weather variables
Calendar variables
Maintenance variables
Methodologies
Model
Statistical regression models
Machine learning models
Neural network models
Hybrid models
Enhancement
Dynamic multi-output framework
Location-conditional concept
Error adjustment strategy
Output
Output variables
Delays
Train process times
Impact of delays
Output dimension
Single variable
Multiple variables
Evaluation
Aspect
Accuracy
Precision
Generalisability
Robustness
Interpretability
Practicality
Dimension
Temporal
Spatial
Train specific
Overall

Scope

Long-term prediction models

Short-term prediction models

Scope

Long-term prediction models

- Study how different factors affect train delays
- Use historical data
- Predict delays several days or months in advance
- For both strategic and tactical train traffic planning

Short-term prediction models

LUND UNIVERSITY

Scope

Long-term prediction models

- Study how different factors affect train delays
- Use historical data
- Predict delays several days or months in advance
- For both strategic and tactical train traffic planning

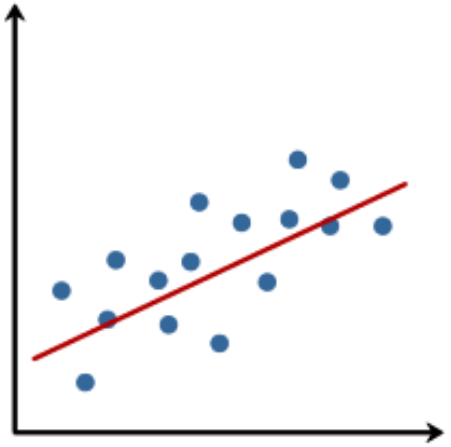
Short-term prediction models

- Focus on making accurate predictions
- Use real-time and historical data
- Predict near-future train delays
- For operational level traffic management

LUND UNIVERSITY

Method

Method

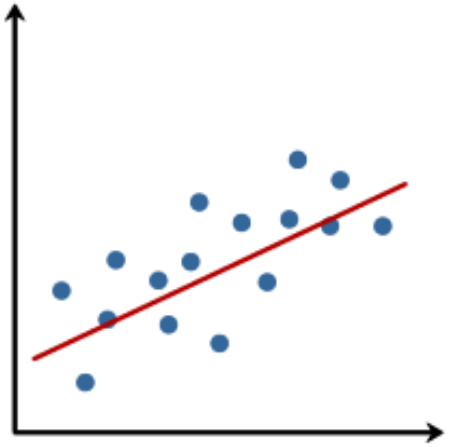


Statistical regression

- It has limitations for modelling complex and non-linear relationships

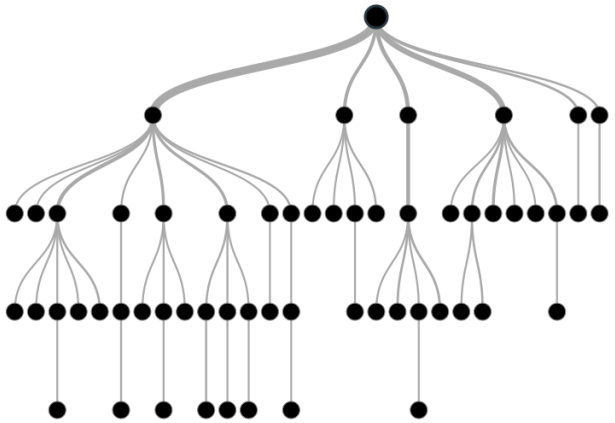
LUND UNIVERSITY

Method



Statistical regression

- It has limitations for modelling complex and non-linear relationships

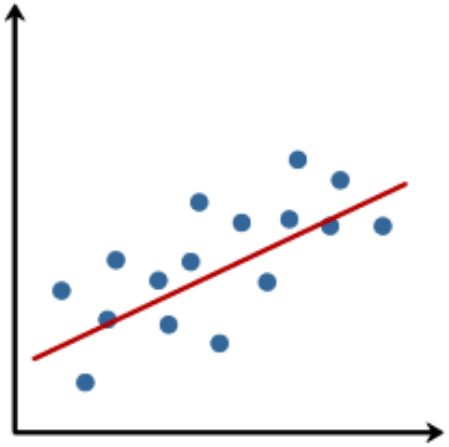


Conventional machine learning

- Less interpretable
- Requires human-engineered spatiotemporal features to capture the spatial and temporal flow patterns of data

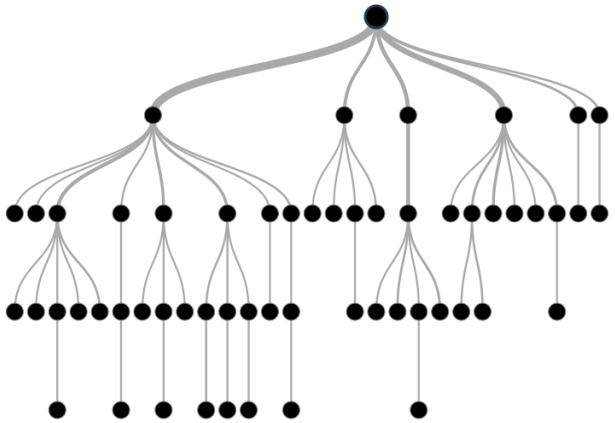
LUND UNIVERSITY

Method



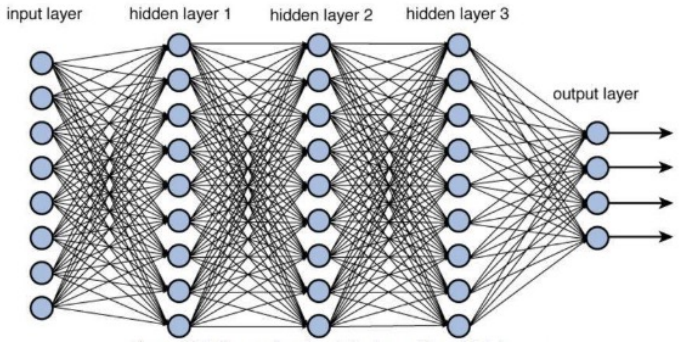
Statistical regression

- It has limitations for modelling complex and non-linear relationships



Conventional machine learning

- Less interpretable
- Requires human-engineered spatiotemporal features to capture the spatial and temporal flow patterns of data

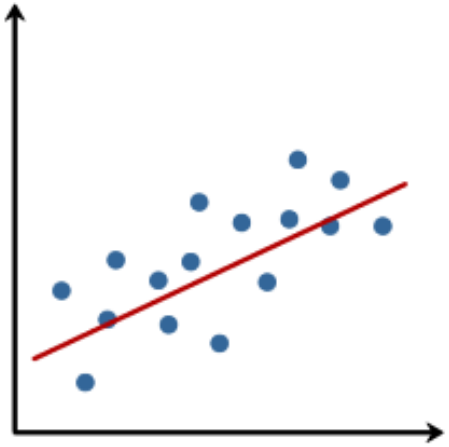


Neural Network

- Automatic learning of spatiotemporal representations from data
- Flexibility to integrate different architectures into hybrid models

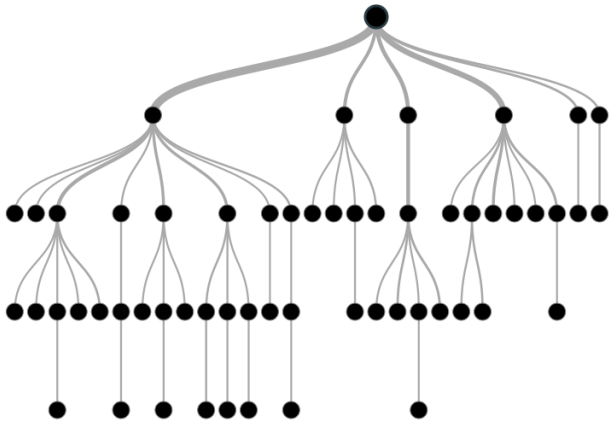
LUND UNIVERSITY

Method



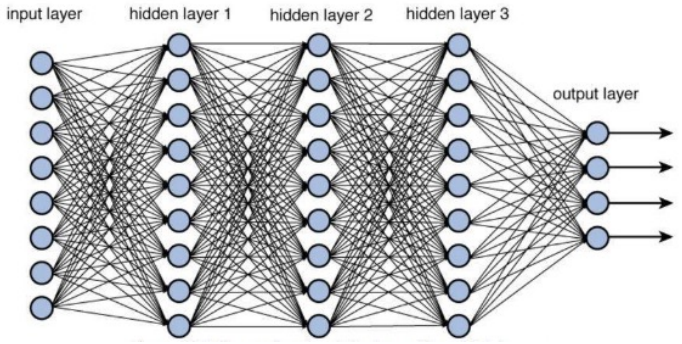
Statistical regression

- It has limitations for modelling complex and non-linear relationships



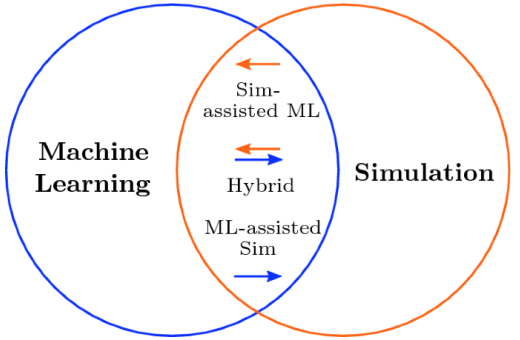
Conventional machine learning

- Less interpretable
- Requires human-engineered spatiotemporal features to capture the spatial and temporal flow patterns of data



Neural Network

- Automatic learning of spatiotemporal representations from data
- Flexibility to integrate different architectures into hybrid models



Hybrid model

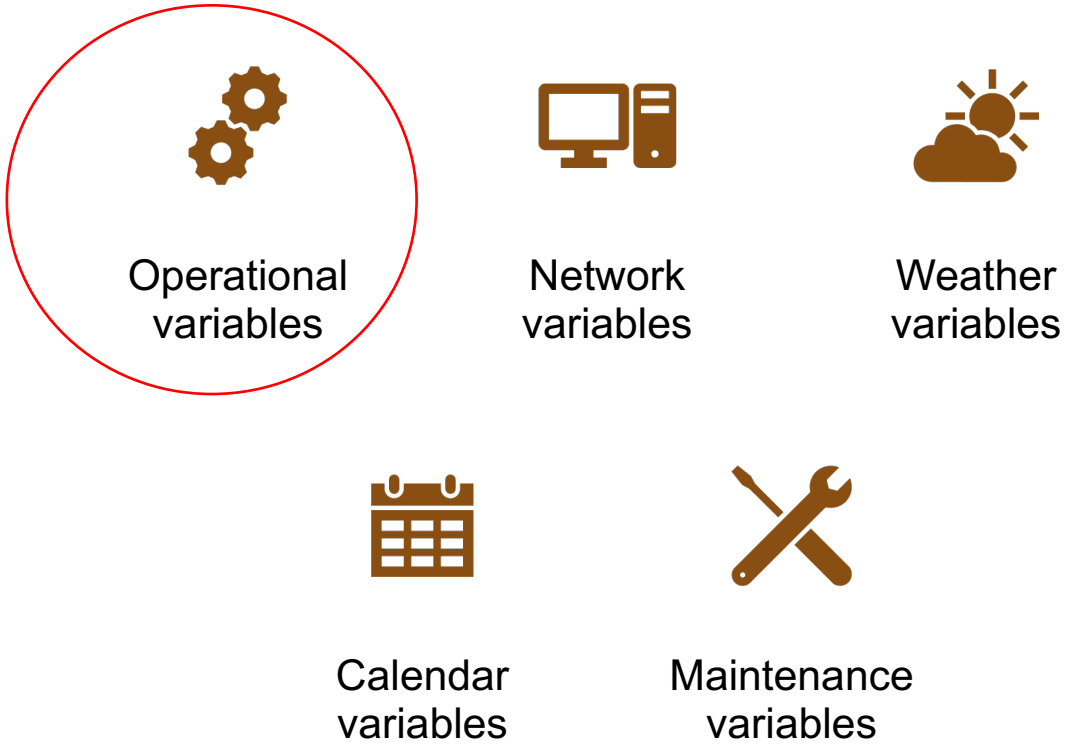
- Multiple base models with uncorrelated prediction errors

LUND UNIVERSITY

Research question 2

How are selected input variables improving the performance of the train delay prediction model?

Input

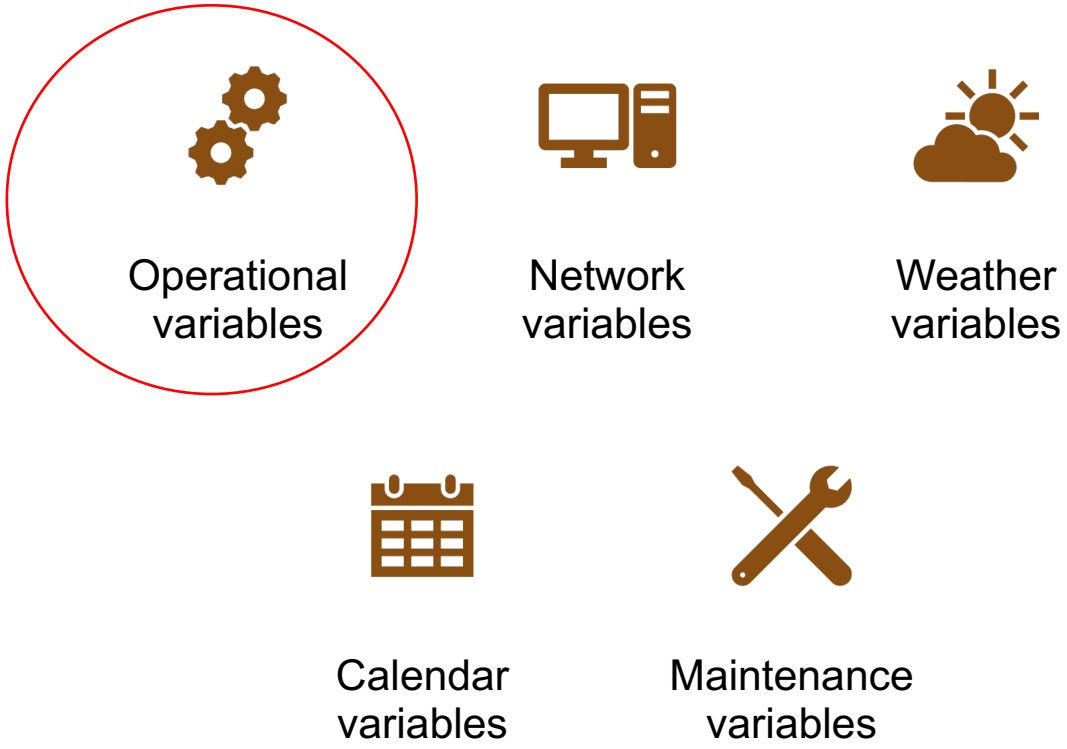


Findings

- The train operation data greatly influences delays.

LUND UNIVERSITY

Input

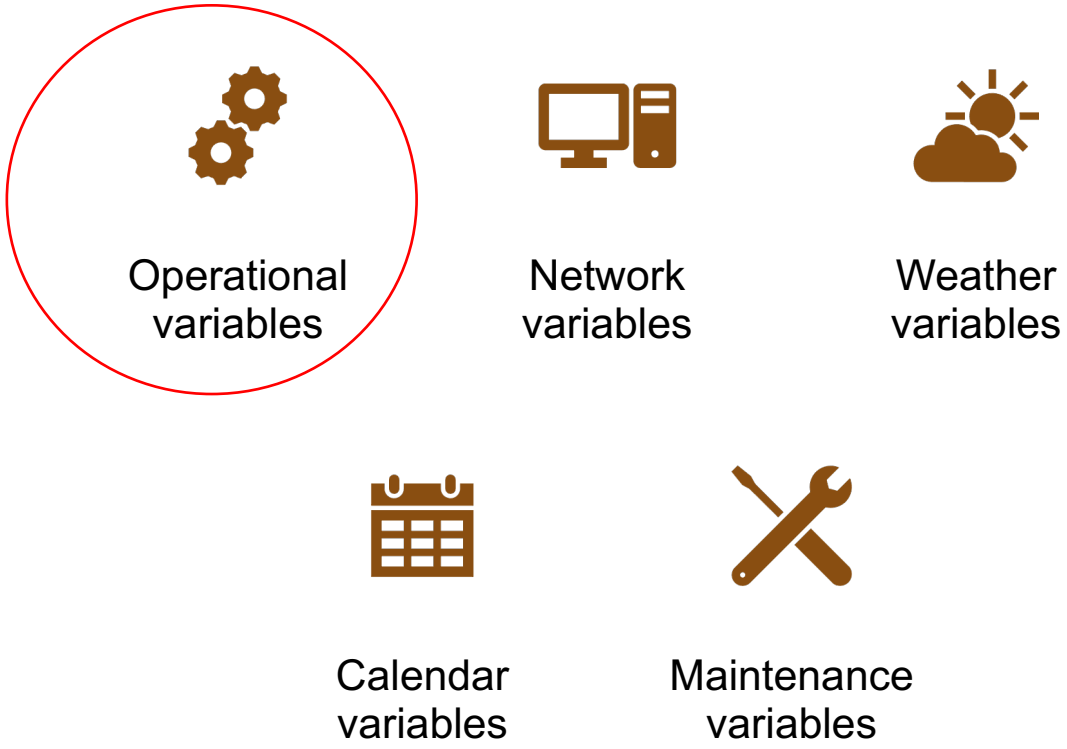


Findings

- The train operation data greatly influences delays.
- Other data adds a layer of adaptability.

LUND UNIVERSITY

Input



Findings

- The train operation data greatly influences delays.
- Other data adds a layer of adaptability.
- The recent observations from nearby stations or trains are important.

LUND UNIVERSITY

Research question 3

What approaches can enhance the train delay prediction model?

Location-conditioned concept

$$\hat{y} = f(X|i)$$

where $\hat{y} = (\hat{t}_{i+1}, \hat{t}_{i+2}, \ldots, \hat{t}_N)$, denotes the predicted train arrival delays at subsequent stations given current station *i*. *X* represents a set of predictor variables encompassing both historical and real-time explanatory factors.

Findings

- Regression models trained conditionally on current train location.

LUND UNIVERSITY

Location-conditioned concept

$$\hat{y} = f(X|i)$$

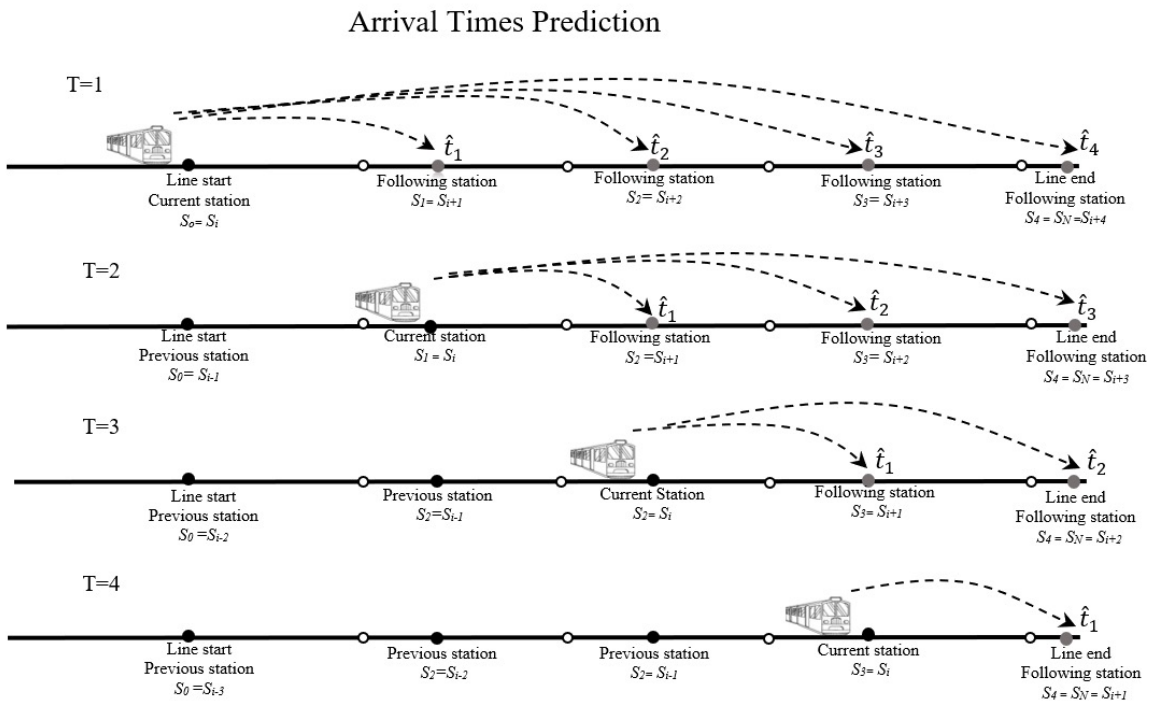
where $\hat{y} = (\hat{t}_{i+1}, \hat{t}_{i+2}, \ldots, \hat{t}_N)$, denotes the predicted train arrival delays at subsequent stations given current station *i*. *X* represents a set of predictor variables encompassing both historical and real-time explanatory factors.

Findings

- Regression models trained conditionally on current train location.
- Considers observable real-time and historical data.

LUND UNIVERSITY

Multi-output framework



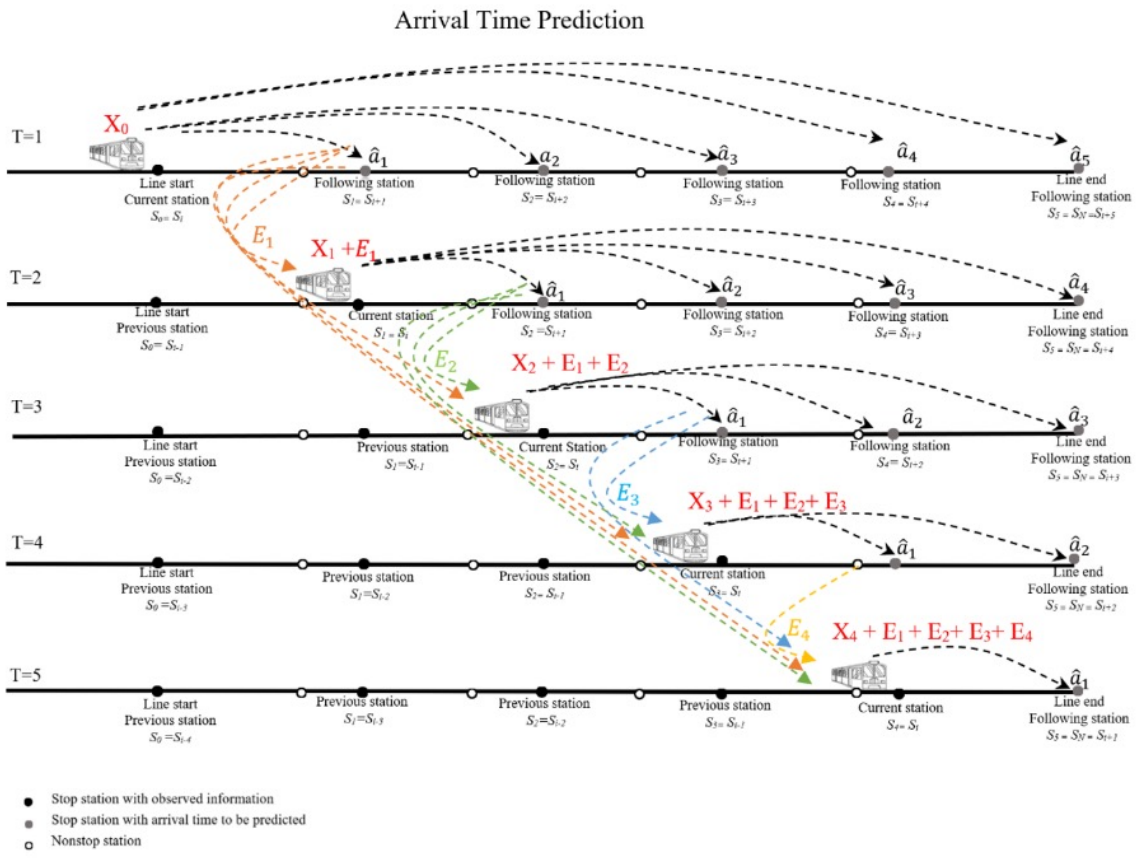
- Stop station with observed information
- Stop station with arrival time to be predicted
- Nonstop station

Findings

- Predict arrival delays for multiple downstream stations at arbitrary times.

LUND UNIVERSITY

Error adjustment strategies



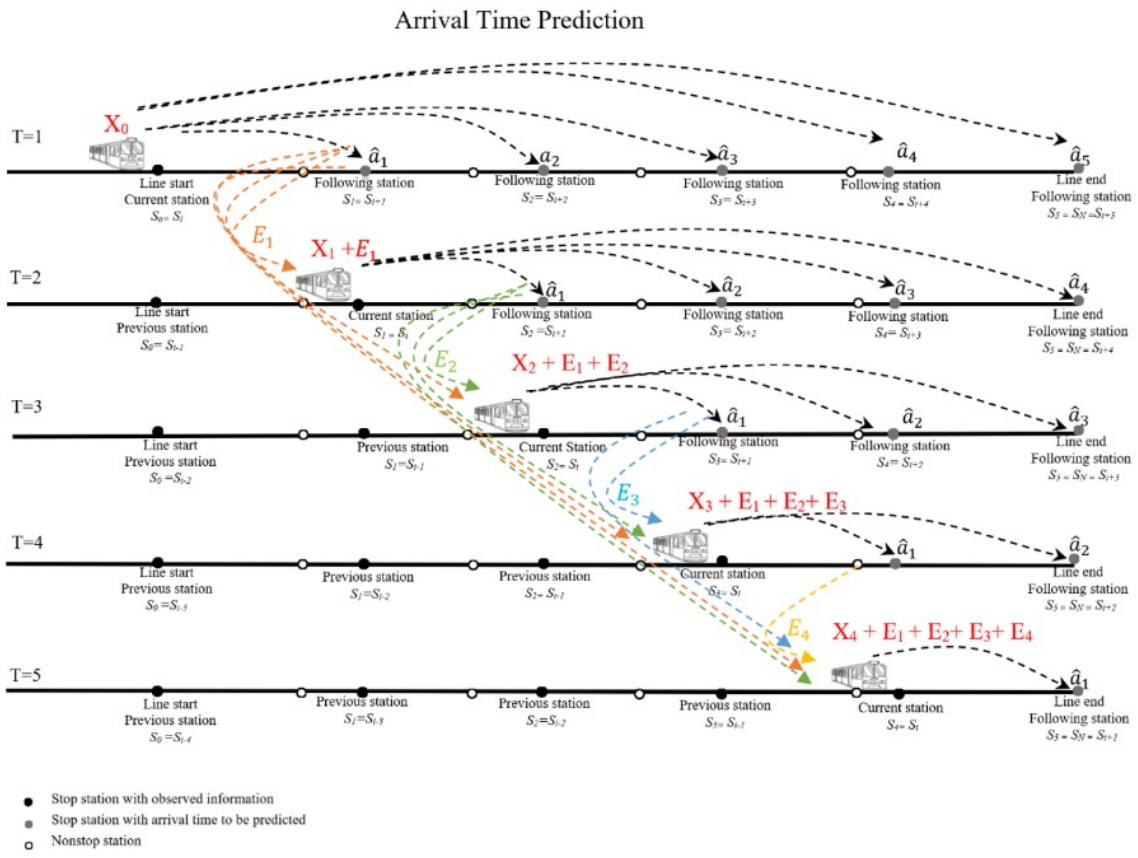
Upstream prediction error correction

Findings

- Use observed information, prediction errors at current and previous stations.

LUND UNIVERSITY

Error adjustment strategies



Upstream prediction error correction

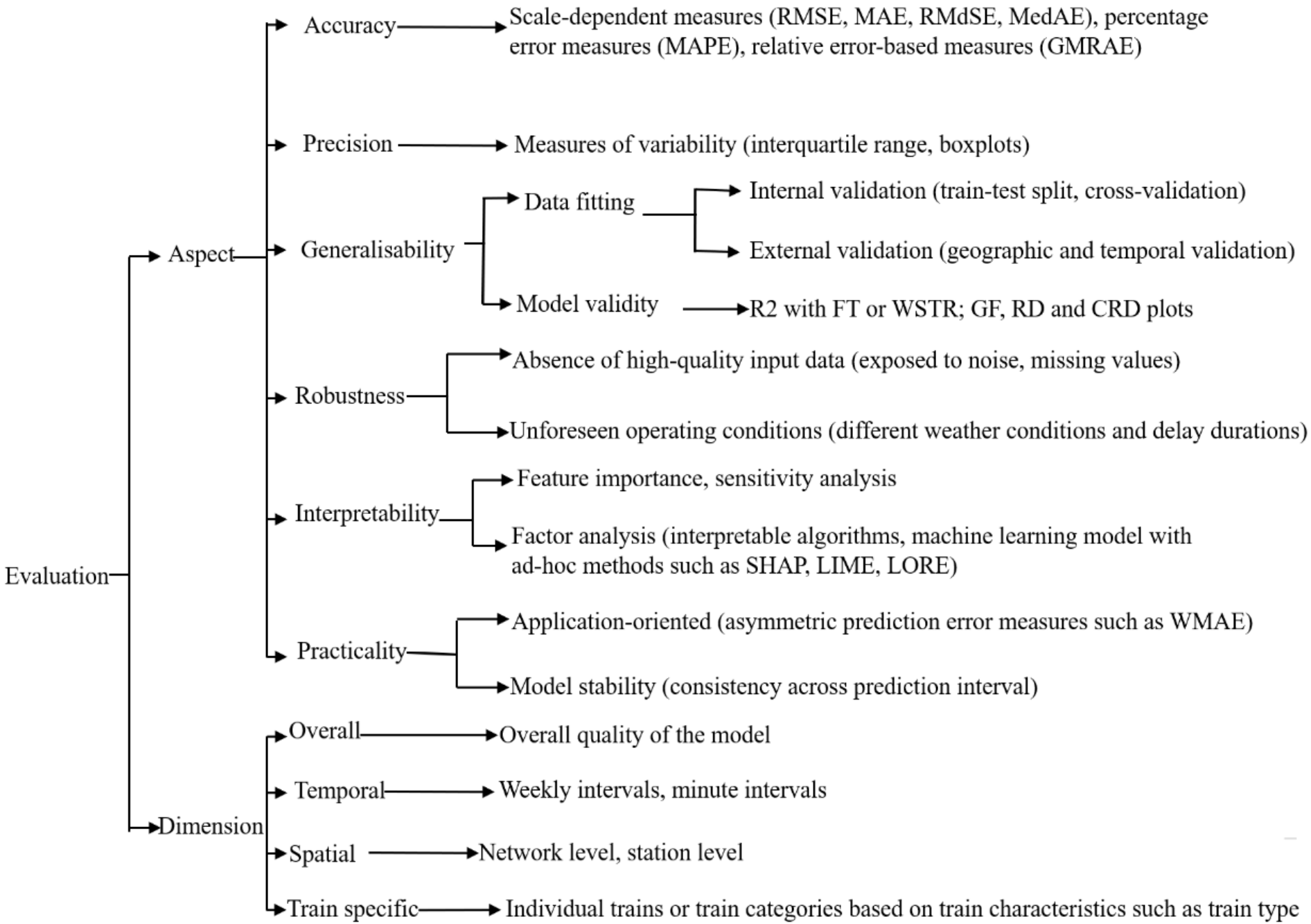
Findings

- Use observed information, prediction errors at current and previous stations.
- Enable the model to constantly adjust itself .

LUND UNIVERSITY

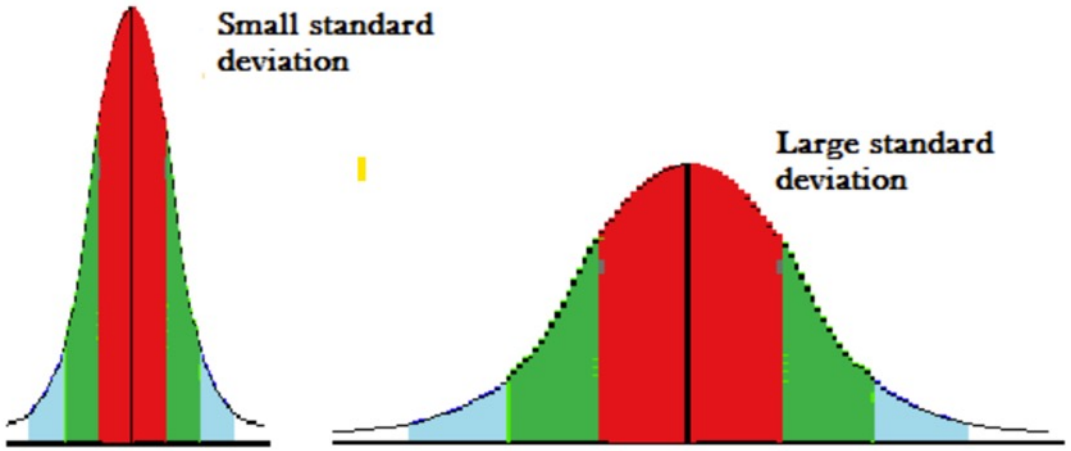
Research question 4

How can train delay prediction models be evaluated?



LUND UNIVERSITY

Precision



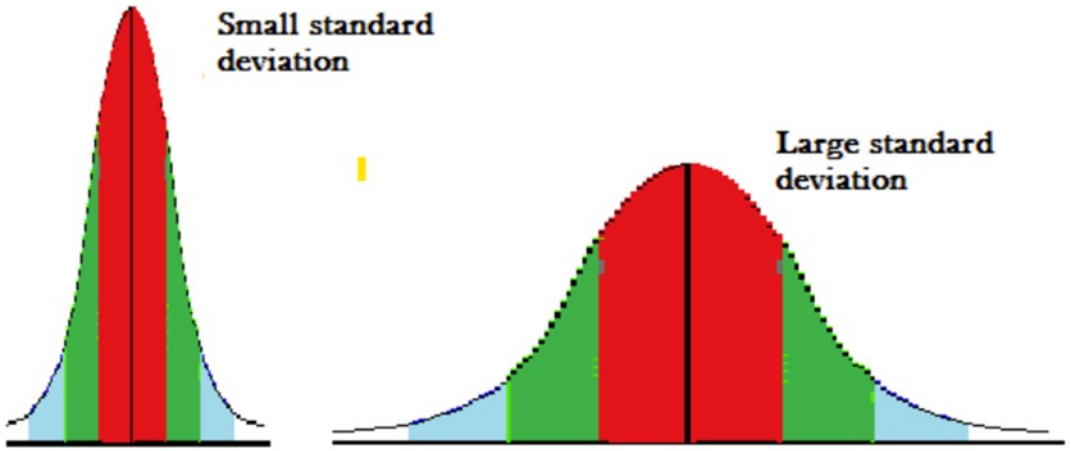
Statistical variance or the spread of data

Findings

- Measures dispersion of prediction error and bias tendencies.

LUND UNIVERSITY

Precision



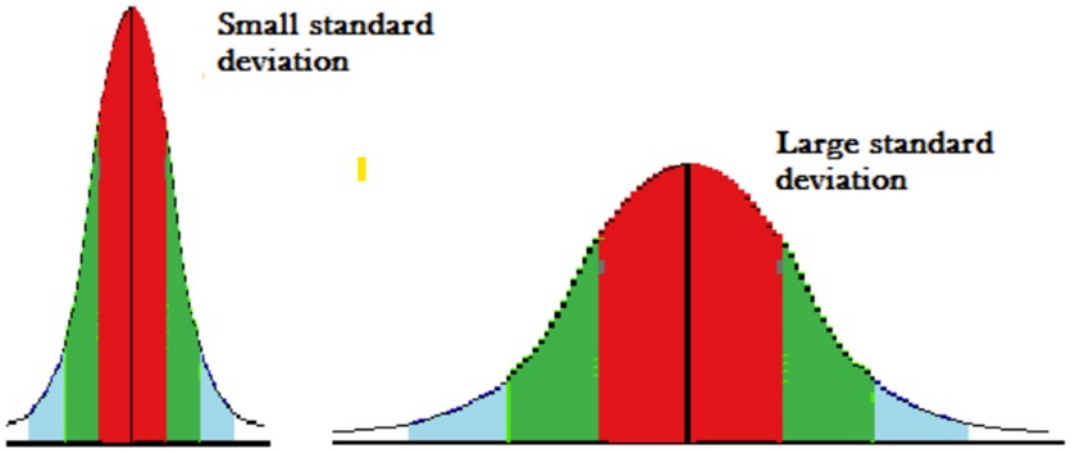
Statistical variance or the spread of data

Findings

- Measures dispersion of prediction error and bias tendencies.
- The narrower ranges mean more reliable predictions.

LUND UNIVERSITY

Precision



Statistical variance or the spread of data

Findings

- Measures dispersion of prediction error and bias tendencies.
- The narrower ranges mean more reliable predictions.
- The interquartile range or boxplot clarify prediction error uncertainty.

LUND UNIVERSITY

Robustness

(a) Invalid inputs

Robustness

(a) Invalid inputs

(b) Challenging environmental conditions

LUND UNIVERSITY

Robustness

Findings

(a) Invalid inputs

(b) Challenging environmental conditions

- Use datasets with realistic representative of real-world application scenarios.

LUND UNIVERSITY

Robustness

(a) Invalid inputs

(b) Challenging environmental conditions

Findings

- Use datasets with realistic representative of real-world application scenarios.
- Prevent purely academic contributions without real-world industrial use.

LUND UNIVERSITY

Practicality

(a) Application-oriented

LUND UNIVERSITY

Practicality

(a) Application-oriented

Findings

- Tolerance for prediction errors varies depending on the model's use case.

LUND UNIVERSITY

Practicality

(a) Application-oriented

Findings

- Tolerance for prediction errors varies depending on the model's use case.
- Measure using asymmetric prediction error measures

LUND UNIVERSITY

Practicality

(a) Application-oriented

Departures			03:38:07
Route	Destination	Stop	Time
1	Blackyard leys	R2	~ 5 Min
4A	Wood farm	R7	~ 20 Min
4	Abingdon	R8	~ 40 min
5	Old woodstock	R5	~ 05:10 Pm
7	Aylesbury	R2	~ 05:45 PM
S1	Thornhill park & ride	R6	~06:00 PM
S3	Seacourt park & ride	R1	~06:20 PM

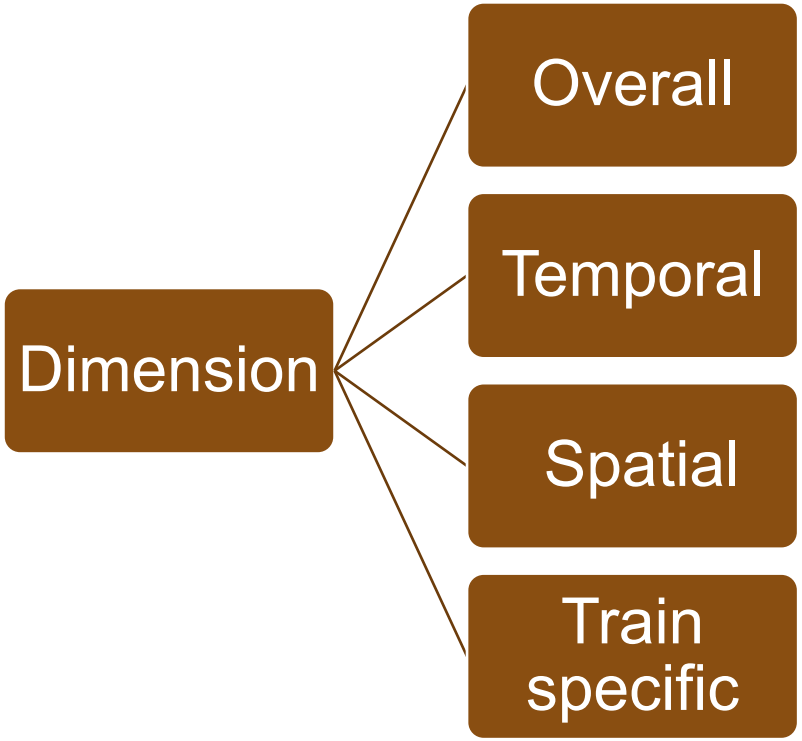
(b) Stability of predictions

Findings

- Tolerance for prediction errors varies depending on the model's use case.
- Measure using asymmetric prediction error measures
- Assesses the consistency of the predictions at each prediction interval .

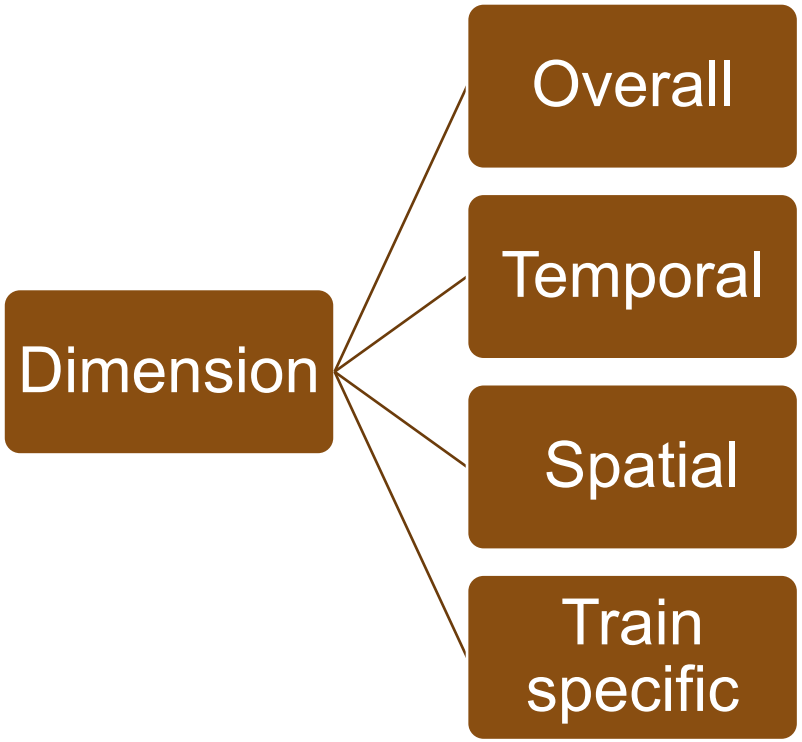
LUND UNIVERSITY

Dimension



LUND UNIVERSITY

Dimension

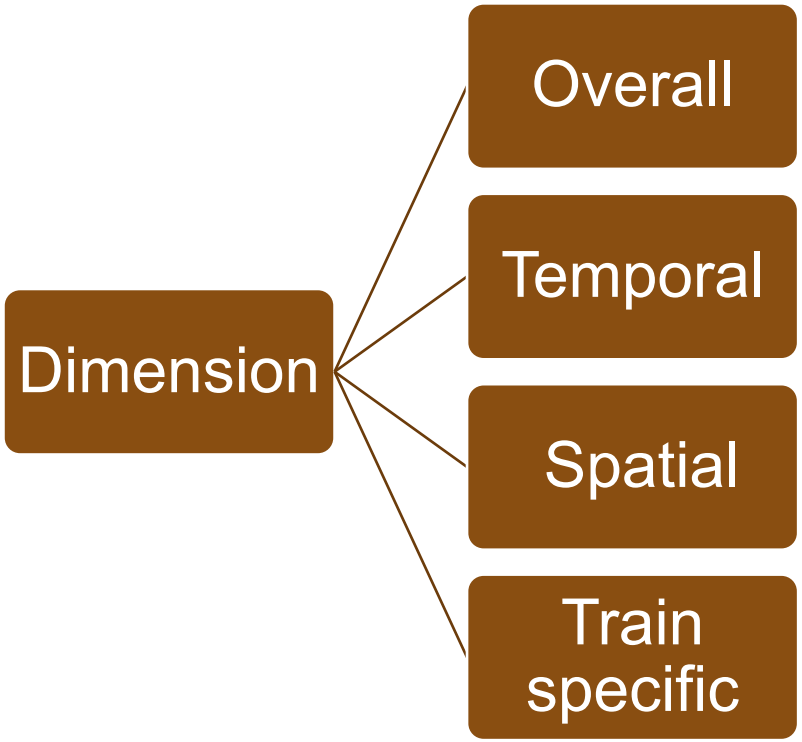


Findings

- Overall performance evaluation provides overview of the model's quality.

LUND UNIVERSITY

Dimension



Findings

- Overall performance evaluation provides overview of the model's quality.
- Detailed evaluations across dimensions uncover underlying performance patterns.

LUND UNIVERSITY

Conclusion & Future research

Conclusion

- Use recent data improves train delay prediction model.
  - Introduced location conditional concepts and error adjustment strategies
  - Generate synthetic train events

LUND UNIVERSITY

Conclusion

- Use recent data improves train delay prediction model.
  - ➢ Introduced location conditional concepts and error adjustment strategies
  - ➢ Generate synthetic train events
- Dynamic multi-output prediction models are crucial for practical applications
  - ➢ Introduced line-level multi-output machine learning models
  - ➢ Network-level prediction models

LUND UNIVERSITY

Conclusion

- Use recent data improves train delay prediction model.
  - ➢Introduced location conditional concepts and error adjustment strategies
  - ➢Generate synthetic train events
- Dynamic multi-output prediction models are crucial for practical applications
  - ➢Introduced line-level multi-output machine learning models
  - ➢Network-level prediction models
- Evaluate models from various aspects and dimensions
  - ➢Established an evaluation framework
  - ➢Conduct comprehensive case studies

LUND UNIVERSITY

Thank you!!!

LUND
UNIVERSITY