Distance estimation in vehicle routing problems

An empirical approach using neural networks and ensemble learning

Jens Mueller

Bocconi University

Master's Thesis Defense
April 22, 2022

Agenda

1 Vehicle Routing Problems (VRPs)

2 Route Distance Estimation
   Relevance
   Literature review

3 Methodology
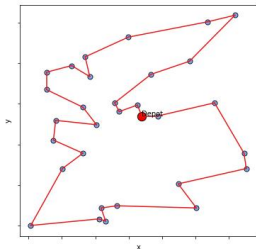   Research design
   Dataset
   Models

4 Results

5 Discussion

Vehicle Routing Problems (VRPs)

- Group of combinatorial optimization problems
- Problem description:
  - **Given:** 1. Set of customers, 2. Fleet of vehicles, 3. Various constraints
  - **Determine:** Feasible routes at minimum cost (typically total distance)
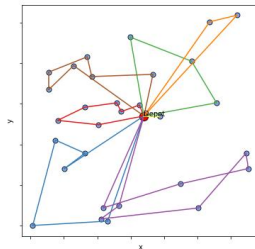- Computationally expensive (NP-hard)
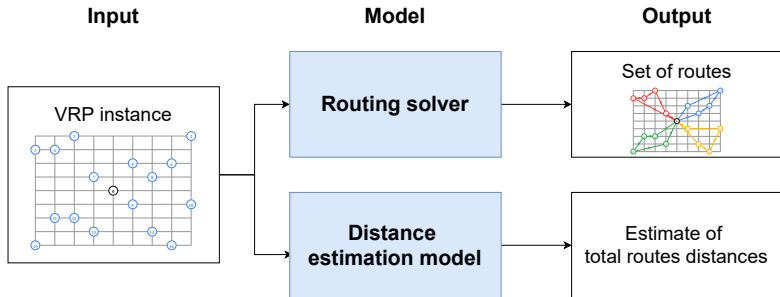

Traveling Salesman Problem (**TSP**)

Capacitated Vehicle Routing Problem (**CVRP**)
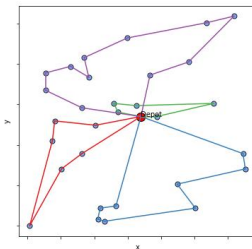
Capacitated VRP with Time Windows (**CVRPTW**)

Route Distance Estimation

Why is distance estimation relevant?

- Integrated routing problems (e.g. location routing)
- Combinatorial auctions
- Managerial decisions

Literature Review

- Mostly linear regression with few predictors and strong assumptions about instance characteristics

- Many papers about the TSP, very few about the CVRPTW

- Current CVRPTW datasets are not well suited for distance estimation.
  - Only four public CVRPTW datasets in VRP-REP
  - Few instances, little variety in instance characteristics, old
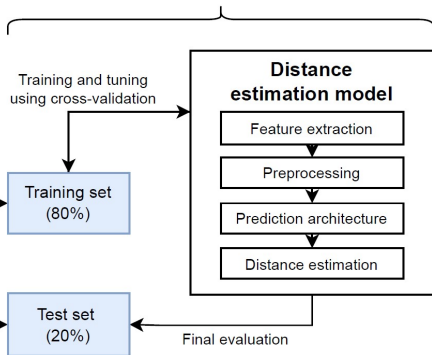
Research objectives

1. A large CVRPTW **dataset** with instance characteristics from a wide variety of distributions

2. New estimation **models** to predict distances more accurately than the linear approaches
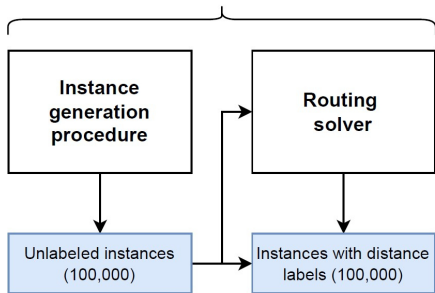
Research Design
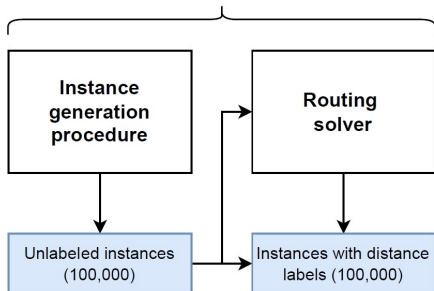
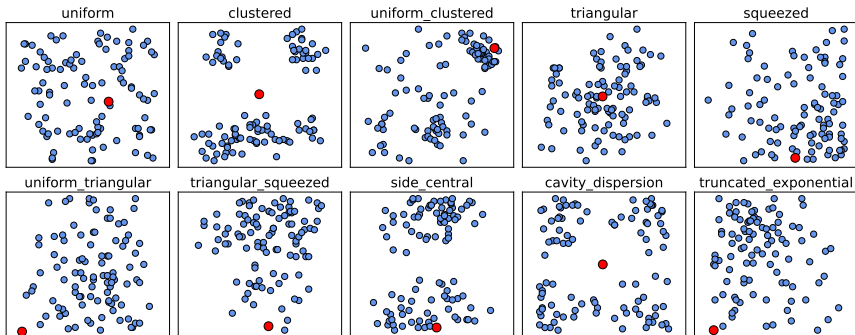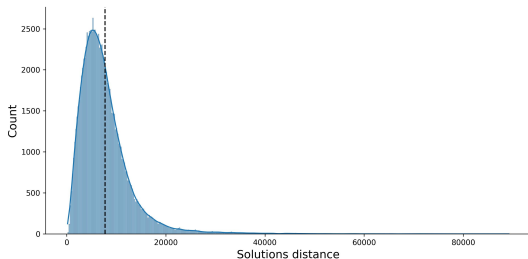# Research Design
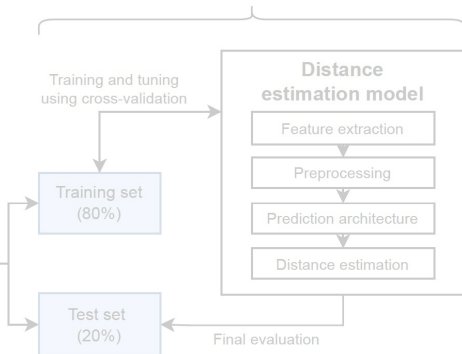
Instance Generation Procedure

- A CVRPTW instance has 12 characteristics.
- Characteristics are sampled from 28 distributions.
- Examples:
  - The number of customers is uniform between 20 and 100.
 - Customer locations follow one of 10 distributions:

Routing Solver

- Meta-heuristics are most common in practice

- Implementation of 16 strategies using OR-Tools
  - Final solver: Path cheapest arc + Guided local search
  - 0.74% to optimal on benchmark by Solomon (1987)

- Full dataset solved on AWS c6g over 5000 hours (3min per instance)
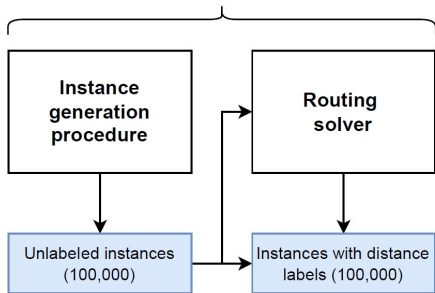

Distribution of solver distances

Log-transformation
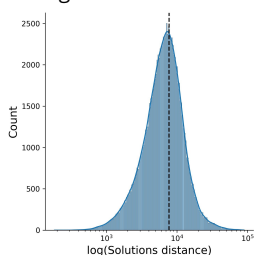
Research Design
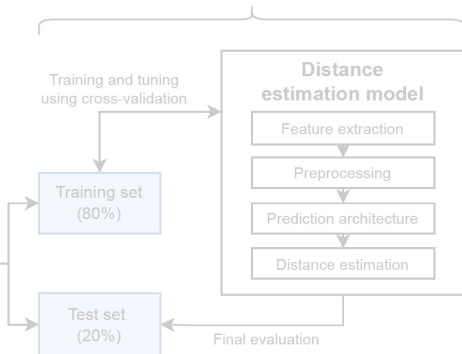
Research Design
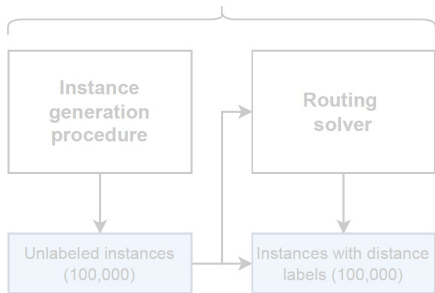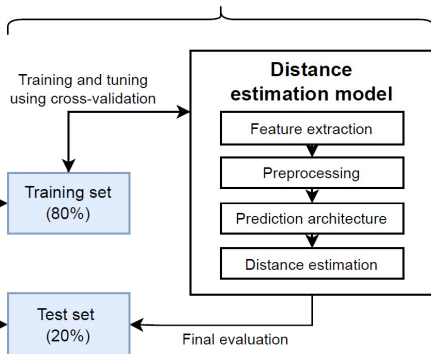
Feature Extraction

- Goal: Capture informative signal from routing instances

- Definition of 45 different features

- Examples:
  - Number of customers
  - Size of the service area
  - Distance from depot to customers
  - Demand coverage
  - Average length of time windows
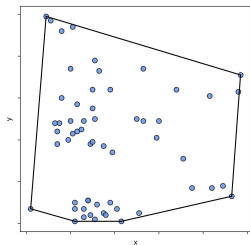
Convex hull

Model Architectures

Baseline models
- Constant model
- Other variants
- Greedy heuristic

Linear regression
- Target transformation
- Feature selection
- Polynomial terms

Multilayer perceptron (MLP)
- Standardization $x' = \frac{x - \mu}{\sigma}$

Ensemble methods
- Random forest
- Gradient boosting
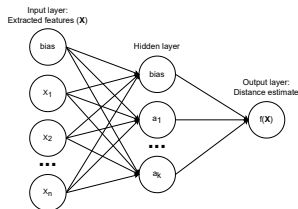
Multilayer perceptron
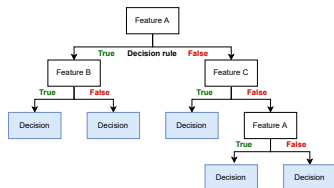
Decision tree

Model Training

Optimization objective is the Root Mean Squared Error:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2}$$

Hyperparameter tuning:

- Random grid search
- Large parameter space
- Iterative refinement
- Over 100 models fitted

Multilayer perceptron training

Results

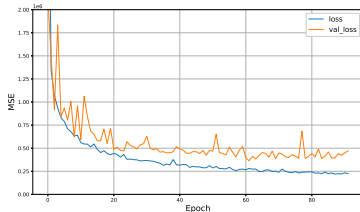| Model	RMSE	MAPE	Prediction time / 1,000 instances
Multilayer-perceptron	636.7	4.48%	8.57s
Gradient boosting	695.9	4.80%	8.54s
Polynomial regression	859.2	5.56%	8.55s
Random forest	1045.0	6.83%	8.55s
Greedy heuristic	1115.6	10.11%	734.14s
Linear regression	1982.1	13.54%	8.53s
Daganzo (1984) - CVRP	2055.6	16.93%	2.85s
Beardwood(1959) - TSP	4498.6	28.49%	2.80s
Constant model	5627.0	78.42%	0.00s

- The new models can predict distances more accurately.
- No significant increase in computation time.
- Routing solver are too slow for distance estimation.

Discussion

Opportunities

- More informed decision-making
- Economic value even for small improvements
- Estimating heuristics is is realistic
- Easily adaptable to other routing variants

Limitations & Future Research

- Bad generalization to larger problems $\rightarrow$ active research field
- Fixed feature vectors $\rightarrow$ graph neural networks
- Slow feature extraction $\rightarrow$ improve code performance
- Real-world data

Conclusion

Topic relevance

- Distance estimation is important in situations that require cost estimates for a large number of instances in short computation time.

Research problems

- Current CVRPTW datasets are not suitable distance estimation.
- Most studies rely on linear models, few predictors, and strong assumptions.

Methodology

- A large CVRPTW dataset with high variance is created and solved.
- Three new distance estimation models are developed.

Results

- The new models achieve better estimates at similar computation time.
- Particularly the MLP and gradient boosting show promising results.
- Further exploration of this approach in richer routing variants is suggested.