# Graph Embeddings for Non-IID Data Feature Representation Learning

QIANG SUN

University of Western Australia

Supervised by

Assoc. Prof. Wei Liu, Dr Du Huynh, Assoc. Prof. Mark Reynolds

Department of Computer Science and Software Engineering

November 22, 2024

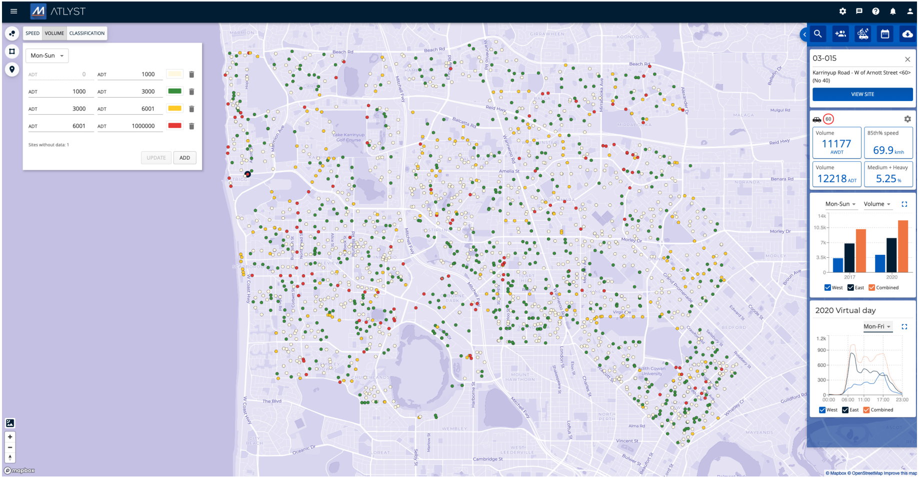
# Contents


1. Motivation and Problem Definition
2. Methodology
3. Dataset and Experiment Design
4. Results and Discussion
5. Conclusion and Future Work

# Motivation: How we handle data now?

| Site name | Year | Description | Asset number | Lat | Lng | ADT | AWDT | AWEDT | 85% speed | Light vehicles(%) | Medium vehicles(%) | Heavy vehicles(%) | Cycle(%) | Motorcycle(%) | Unclassifiable(%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 05-013 | 2021 | Cobb Street - E of Calais Road <50> (No 77) | | -31.907133 | 115.769148 | 875.00 | 899.00 | 797.00 | 54.11 | 90.66 | 7.56 | 0.11 | 0.91 | 0.69 | 0.07 |
| 05-062 | 2021 | Westview Street - N of Scarborough Beach Road <50> (No 151) | | -31.893093 | 115.774083 | 857.00 | 879.00 | 790.00 | 36.68 | 93.76 | 3.57 | 0.07 | 0.94 | 0.57 | 1.09 |
| 05-014 | 2021 | Cobb Street - E of Cornelian Street <50> (No 122) | | -31.907387 | 115.774543 | 3617.00 | 3838.00 | 2954.00 | 54.50 | 94.53 | 4.00 | 0.12 | 0.70 | 0.62 | 0.03 |
| 07-030 | 2021 | Maree Street - W of Blissett Way <50> (No 26) | | -31.851122 | 115.817148 | 223.00 | 234.00 | 203.00 | 54.40 | 94.95 | 3.51 | 0.23 | 0.84 | 0.47 | 0.00 |
| 08-027 | 2021 | Swiftlet Way - N of Willowbank Entrance <50> (No 4) | | -31.866735 | 115.791622 | 639.00 | 648.00 | 604.00 | 38.02 | 92.95 | 5.42 | 0.12 | 0.60 | 0.91 | 0.00 |
| 06-026 | 2021 | Monyash Road - S of Wessex Street <50> (No 28) | | -31.853747 | 115.790568 | 278.00 | 266.00 | 322.00 | 49.39 | 90.82 | 7.17 | 0.35 | 0.99 | 0.62 | 0.05 |
| 05-054 | 2021 | Stanley Street - N of Ventnor Street <50> (No 100B) | | -31.902667 | 115.762312 | 431.00 | 437.00 | 416.00 | 47.02 | 95.20 | 3.28 | 0.00 | 0.32 | 1.08 | 0.12 |
| 06-006 | 2021 | Camelot Street - E of Duffy Road <50> (No 7) | | -31.853608 | 115.794553 | 1119.00 | 1162.00 | 980.00 | 45.68 | 93.24 | 5.25 | 0.70 | 0.33 | 0.45 | 0.03 |
| 06-001 | 2021 | Almadine Drive - E of Marmion Avenue <50> (No 52) | | -31.853632 | 115.768778 | 2011.00 | 2387.00 | 873.00 | 54.61 | 94.29 | 4.33 | 0.31 | 0.48 | 0.51 | 0.08 |
| 05-028 | 2021 | Duke Street - N of Brighton Road <50> (No 193) | | -31.896198 | 115.771465 | 7359.00 | 7728.00 | 6266.00 | 58.00 | 92.68 | 6.55 | 0.26 | 0.07 | 0.42 | 0.03 |
| 05-022 | 2021 | Dover Road - N of Ventnor Street <50> (No 45) | | -31.902717 | 115.765752 | 407.00 | 416.00 | 367.00 | 53.71 | 94.05 | 3.82 | 0.06 | 0.69 | 1.38 | 0.00 |
| 05-050 | 2021 | Sackville Terrace - W of Abbett Street <60> (No 77 Abbett) | | -31.887698 | 115.768812 | 6965.00 | 6801.00 | 7455.00 | 47.81 | 92.40 | 5.90 | 0.21 | 0.53 | 0.87 | 0.10 |
| 06-043 | 2021 | Whittington Avenue - W of Godecke Rise <50> (No 12) | | -31.847642 | 115.769412 | 1169.00 | 1174.00 | 1151.00 | 55.58 | 95.72 | 2.93 | 0.16 | 0.42 | 0.70 | 0.07 |
| 08-003 | 2021 | Balcatta Road West - E of Careniup Avenue <50> (70m East) | | -31.85932 | 115.796223 | 564.00 | 582.00 | 504.00 | 57.71 | 90.62 | 7.52 | 0.51 | 1.01 | 0.29 | 0.04 |
| 06-042 | 2021 | Whittington Avenue - E of Bradbourne Drive <50> (No 54) | | -31.848895 | 115.765343 | 1015.00 | 1020.00 | 999.00 | 56.30 | 94.49 | 3.89 | 0.25 | 0.46 | 0.88 | 0.02 |
| 18-006 | 2021 | Dolomite Court - E of Silkwood Turn <50> (No 29) | | -31.921695 | 115.79033 | 858.00 | 969.00 | 493.00 | 51.70 | 94.67 | 3.71 | 0.06 | 1.20 | 0.29 | 0.06 |
| 06-014 | 2021 | Edlaston Road - E of Kersey Way E <50> (No 33) | | -31.846663 | 115.775163 | 698.00 | 705.00 | 678.00 | 58.90 | 94.66 | 3.85 | 0.13 | 0.68 | 0.66 | 0.04 |
| 06-016 | 2021 | Everingham Street - S of Osmaston Road <50> (No 76) | | -31.851858 | 115.777517 | 2312.00 | 2838.00 | 727.00 | 50.51 | 94.91 | 3.90 | 0.14 | 0.26 | 0.41 | 0.38 |
| 27-028 | 2021 | Fourth Avenue - W of John Street <50> (No 70) | | -31.926933 | 115.880983 | 1314.00 | 1423.00 | 1145.00 | 54.22 | 94.52 | 3.34 | 0.08 | 1.28 | 0.74 | 0.04 |
| 28-025 | 2021 | Harcourt Street -  E of Beaufort Street <50> (No 13) | | -31.92012 | 115.889367 | 211.00 | 180.00 | 214.00 | 46.30 | 90.56 | 5.41 | 0.32 | 2.42 | 1.13 | 0.16 |
| 28-078 | 2021 | Dundas Road -  E of Arthur Street <50> (No 11) | | -31.919552 | 115.886608 | 1505.00 | 1526.00 | 1450.00 | 52.49 | 94.54 | 2.98 | 0.07 | 1.62 | 0.54 | 0.25 |
| 28-050 | 2021 | Wood Street -  W of Beaufort Street <50> (No 14) | | -31.916963 | 115.889622 | 2095.00 | 2195.00 | 1824.00 | 49.00 | 94.90 | 3.03 | 0.17 | 1.14 | 0.62 | 0.13 |
| 06-013 | 2021 | Edlaston Road - W of Alvaston Drive <50> (No 60) | | -31.846517 | 115.779003 | 1045.00 | 1058.00 | 1009.00 | 52.42 | 94.78 | 4.54 | 0.13 | 0.07 | 0.45 | 0.02 |
| 18-046 | 2021 | Campus Way - W of Pearson Street <50> (No 6) | | -31.925042 | 115.793828 | 318.00 | 326.00 | 283.00 | 44.21 | 94.63 | 3.87 | 0.15 | 0.38 | 0.70 | 0.27 |
| 28-074 | 2021 | Beaufort Street -  S of Eighth Avenue <50> (HN 851 (Hair Studio)) | | -31.922678 | 115.8839 | 22457.00 | 22692.00 | 21826.00 | 50.90 | 93.37 | 5.26 | 0.20 | 0.45 | 0.68 | 0.05 |
| 28-080 | 2021 | Normanby Road -  W of Beaufort Street <50> (No 6) | | -31.919038 | 115.887795 | 265.00 | 279.00 | 227.00 | 45.22 | 92.41 | 3.54 | 0.06 | 3.07 | 0.81 | 0.12 |
| 07-027 | 2021 | Hornet Street - S of Eglinton Crescent <50> (No 4) | | -31.853987 | 115.809802 | 619.00 | 630.00 | 577.00 | 44.89 | 92.55 | 6.19 | 0.21 | 0.35 | 0.56 | 0.14 |
| 06-031 | 2021 | Osmaston Road - S of Edlaston Road <50> (No 47) | | -31.849087 | 115.772533 | 1848.00 | 2233.00 | 690.00 | 53.10 | 94.26 | 4.01 | 0.15 | 0.97 | 0.50 | 0.11 |
| 05-006 | 2021 | Brighton Road - E of Hinderwell Street <50> (No 101) | | -31.898005 | 115.767125 | 8918.00 | 8891.00 | 9047.00 | 54.22 | 94.77 | 4.03 | 0.08 | 0.30 | 0.77 | 0.05 |
| 05-009 | 2021 | Burniston Street - N of Scarborough Beach Road <50> (No 145) | | -31.892408 | 115.770427 | 877.00 | 928.00 | 719.00 | 51.01 | 94.28 | 3.98 | 0.36 | 0.41 | 0.89 | 0.09 |
| 07-019 | 2021 | Eglinton Crescent - W of Erindale Road <50> (No 147) | | -31.853707 | 115.8118 | 2432.00 | 2820.00 | 1262.00 | 59.69 | 92.68 | 6.22 | 0.25 | 0.27 | 0.53 | 0.06 |
| 30-073 | 2021 | The Strand - E of McLean Street <50> (No 332) | | -31.896713 | 115.878838 | 2482.00 | 2533.00 | 2320.00 | 59.51 | 94.36 | 4.58 | 0.23 | 0.49 | 0.26 | 0.08 |
| 18-043 | 2021 | Mountainbell Road - N of Cromarty Road <50> (20m North) | | -31.927283 | 115.793548 | 452.00 | 462.00 | 398.00 | 34.60 | 82.49 | 11.26 | 0.03 | 0.97 | 0.57 | 4.69 |
| 18-044 | 2021 | Parrotbush Road - N of Cromarty Road <50> (No 4) | | -31.927255 | 115.792547 | 839.00 | 878.00 | 699.00 | 40.21 | 84.73 | 6.67 | 0.10 | 0.85 | 0.38 | 7.27 |
| 28-075 | 2021 | Nelson Street -  N of York Street <50> (No 20) | | -31.918648 | 115.892252 | 905.00 | 932.00 | 832.00 | 53.10 | 94.32 | 3.63 | 0.02 | 1.38 | 0.63 | 0.02 |
| 28-077 | 2021 | Central Avenue -  W of Beaufort Street <50> (No 145) | | -31.924177 | 115.880942 | 13448.00 | 14180.00 | 11595.00 | 60.52 | 95.61 | 3.33 | 0.17 | 0.25 | 0.60 | 0.04 |

# Motivation: Visualization of the structural information



# Motivation: What's Next?

It looks cool, but:

- How to make machines represent, interpret and leverage structured information effectively?

# Problem definition

- How to encode structural information of a network into a machine learning problem?
- Whether the structural enhanced feature space improve classification performance?
- Knowledge graph embedding compares with homogenous network embedding in feature encoding?
- We use a traffic network where spatial information is naturally present.

# Methodology: The I.I.D (Independent Identical Distribution) Assumption

| Models | Rows/Observations | Columns/Features |
|---|---|---|
| Logistic Regression | Assume independent | Assume independent |
| Naïve Bayes | Assume independent | Assume independent |
| SVM | Assume independent | Assume independent |
| DT/RF | Assume independent | Assume independent |
| kNN | No assumption | No assumption |

Liu et al, IJCNN 2014

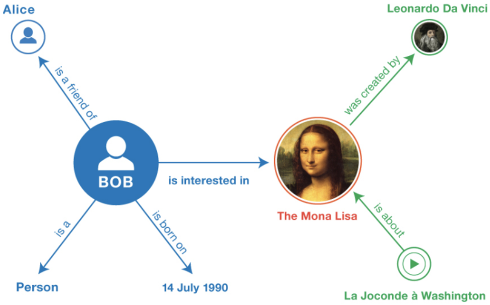
# Methodology: Graph


A simple graph

- Nodes are of the same type.
- Edges can be directed or weighted.

https://web.stanford.edu/class/cs520/2020/notes/What_is_a_Knowledge_Graph.html

# Methodology: Knowledge Graph (KG)

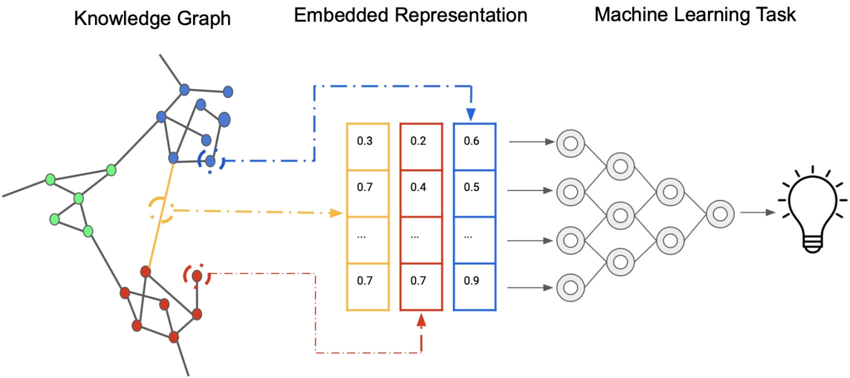



A simple knowledge graph

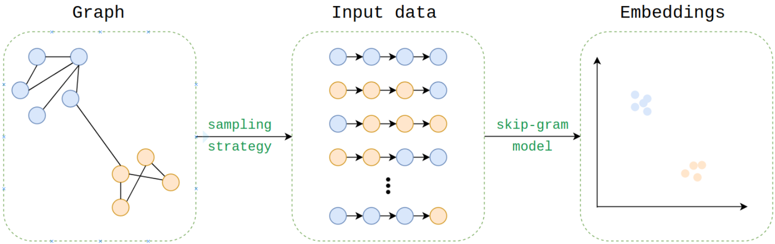
- Nodes and edges all have types.
- Ontology required.

# Methodology: Graph and Knowledge Graph

## Embedding Workflow



# Methodology: Graph Embedding with Node2vec

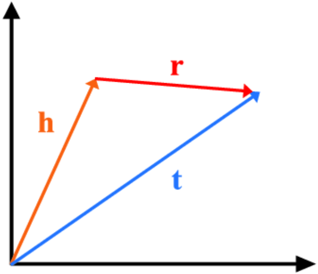



node2vec embedding process

- Biased random walk, which explores neighborhoods in BFS as well as DFS fashion, generate sequences as input.
- Capture homophily and structural equivalence.

https://snap.stanford.edu/node2vec/

# Methodology: Knowledge Graph Embedding with TransE




$$h + r = t$$

https://en.wikipedia.org/wiki/Knowledge_graph_embedding
Bordes et al, Advances in Neural Information Processing Systems, 2013

# Methodology: How we handle spatial datasets?

1. **Graph and Knowledge Graph Construction**: A road network graph and a traffic knowledge graph.
2. **Representation learning**: Compute the embeddings by node2vec, TransE respectively.
3. **Machine Learning Tasks**: Use embeddings as input, apply SVM, kNN and RF.

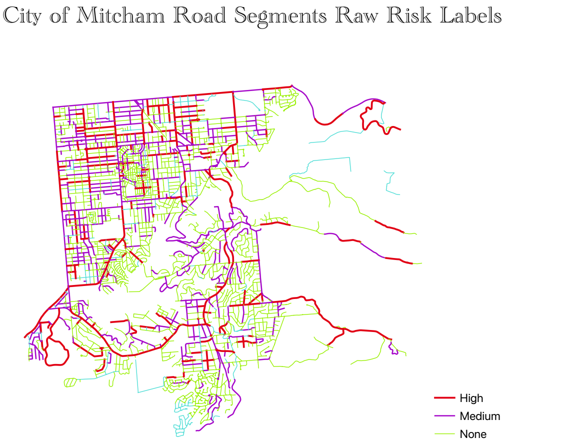
# Dataset

The road network is composed of 2287 road segments, for each road segment, we selected the following 8 features:

- a unique *asset id* to identify road segments
- *speed limit*, eg: 40km/h, 50km/h, 60km/h
- difference between speed limit and 85*th* %*ile* speed
- average daily *AM peak hour traffic volume*
- average daily *PM peak hour traffic volume*
- total *number of rear end hit* accidents since 2016
- total *number of crashes* since 2016
- total *number of casualties* since 2016

1 for asset management, 2 for speed, 2 for volume, 3 for risk

# Dataset: Overview



# Graph and Knowledge Graph Construction

1. **Graph and Knowledge Graph Construction**: A road network graph and a traffic knowledge graph.
2. **Representation learning**: Compute the embeddings by node2vec, TransE respectively.
3. **Machine Learning Tasks**: Use embeddings as input, apply SVM, kNN and RF.

# Graph and Knowledge Graph Construction (cont.)

With the help of Neo4j[1], we build a simple road network graph, which contains the following:

| Name | Type | Comments | Numbers |
|---|---|---|---|
| **RoadSegment** | Node | Use asset id to distinguish nodes | 2287 |
| **Link** | Edge | Link connected RoadSegment | 3772 |

To build a knowledge graph upon this:

- Keep the road segment nodes.
- Discrete attributes to obtain nodes for *speed/volume/risk*.

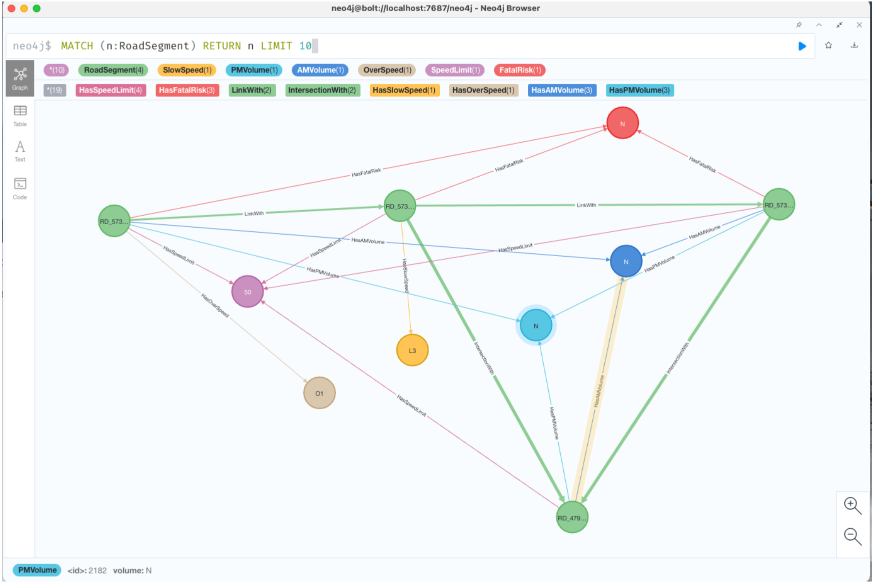
[1] https://neo4j.com/

# Graph and Knowledge Graph Construction (cont.)

Same, with Neo4j, the summary of traffic knowledge graph we built as following:

| Name | Type | Explain | Numbers |
|---|---|---|---|
| **RoadSegment** | Node | Use ASSET_ID to distinguish nodes | 2287 |
| **FatalRisk** | Node | Enumerate, H/M/N, stands for risk level | 3 |
| **RearEndHit** | Node | RearEndHit occurred | 1 |
| **OverSpeed** | Node | Enumerate, $O_1/O_2/O_3/O_4/O_5$ | 5 |
| **SlowSpeed** | Node | Enumerate, $L_1/L_2/L_3/L_4/L_5$ | 5 |
| **SpeedLimit** | Node | Enumerate, 15/25/40/50/60/70 (km/h) | 6 |
| **AMVolume** | Node | Enumerate, H/M/L/N | 4 |
| **PMVolume** | Node | Enumerate, H/M/L/N | 4 |
| **LinkWith** | Edge | Link connected RoadSegment | 1086 |
| **IntersectionWith** | Edge | The link between roadsegments is intersection | 2686 |
| **HasFatalRisk** | Edge | Link RoadSegment with FatalRisk nodes | 2287 |
| **HasRearEndHit** | Edge | Link RearEndHit with RoadSegment | 117 |
| **HasOverSpeed** | Edge | Link OverSpeed nodes with RoadSegment | 423 |
| **HasSlowSpeed** | Edge | Link SlowSpeed nodes with RoadSegment | 548 |
| **HasSpeedLimit** | Edge | Link SpeedLimit nodes with RoadSegment | 2287 |
| **HasAMVolume** | Edge | Link AMVolume nodes with RoadSegment | 974 |
| **HasPMVolume** | Edge | Link PMVolume nodes with RoadSegment | 974 |

# Graph and Knowledge Graph Construction (cont.)



# Representation Learning

1. **Graph and Knowledge Graph Construction**: A road network graph and a traffic knowledge graph.
2. **Representation learning**: Compute the embeddings by node2vec, TransE respectively.
3. **Machine Learning Tasks**: Use embeddings as input, apply SVM, kNN and RF.

# Representation Learning: Summary of Results

Using *NetworkX*[2], *node2vec*[3], *pykg2vec*[4], we managed to train and compute the embeddings for both the traffic knowledge graph and the road network graph.

- a 50-dimensions vector for each road segment in the road network graph via node2vec.
- a 58-dimensions vector for each road segment in the traffic KG via TransE.

---

[2] https://networkx.org/

[3] https://snap.stanford.edu/node2vec/

[4] https://pykg2vec.readthedocs.io/

# Machine Learning Tasks (cont.)

1. **Graph and Knowledge Graph Construction**: A road network graph and a traffic knowledge graph.
2. **Representation learning**: Compute the embeddings by node2vec, TransE respectively.
3. **Machine Learning Tasks**: Use embeddings as input, apply SVM, kNN and RF.

# Machine Learning Tasks (cont.)

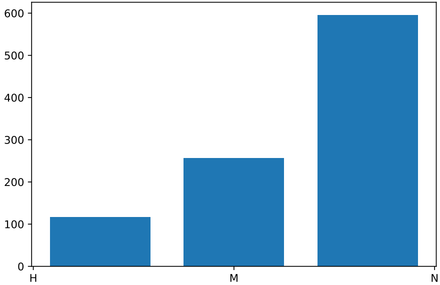
- For our classification task, our target variable is the *risk level*.
- We used SVM, kNN and RF to train the classification models.
- For traditional models, we have 5 input features: *speed limit*, *difference between speed limit and 85th %ile*, *AM peak volume*, *PM peak volume*, *rear end hit number*.
- Only 969 road segments without any missing features, having all the 5 features.
- 1316 have no speed data, and 1313 have no traffic volume data, due to traffic survey limitation.

# Machine Learning Tasks: Input sets for models

We want to see the performance differences when we feed different input sets into models.

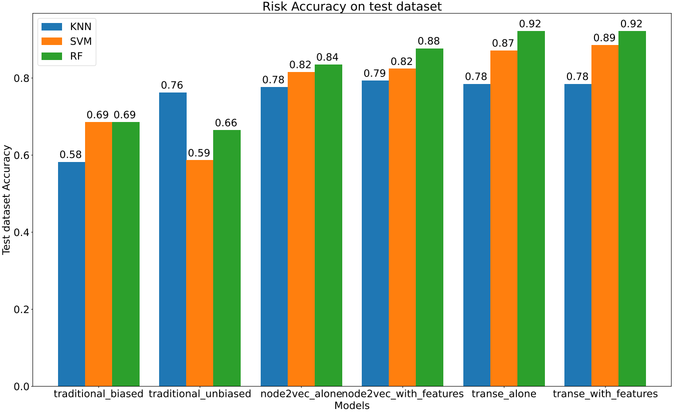
| Input | Reference ID |
|---|---:|
| five features | traditional_biased/unbiased |
| 50-dim vectors (node2vec) | node2vec_alone |
| 50-dim vectors (node2vec) + five features | node2vec_with_features |
| 58-dim vectors (TransE) | transe_alone |
| 58-dim vectors (TransE) + five features | transe_with_features |

# Machine Learning Tasks: Oversampling for highly imbalanced data



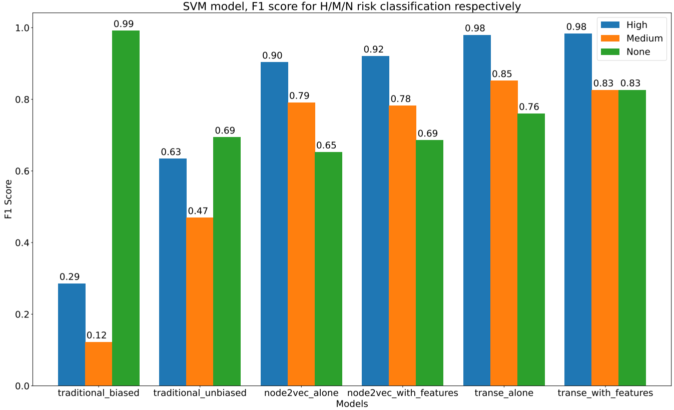

Imbalanced risk level distribution

# Results: Risk classification



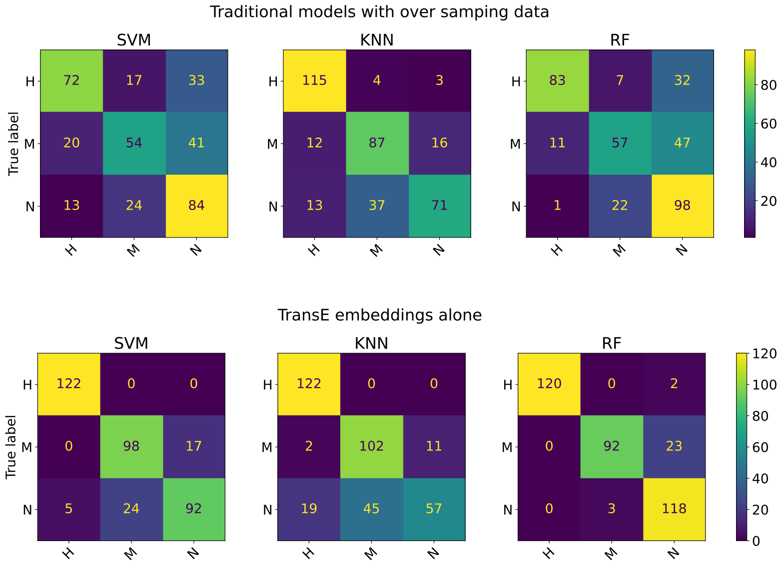

Risk classification accuracy for train and test datasets, per input set, per model

# Results: Risk classification

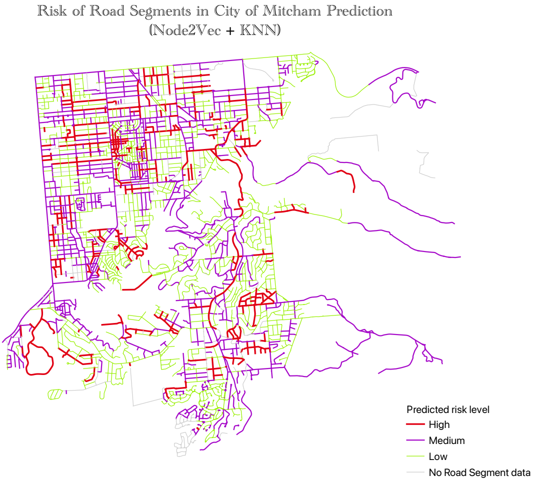



Risk classification f1 score for each input set, per risk level

# Results: Risk classification confusion matrix

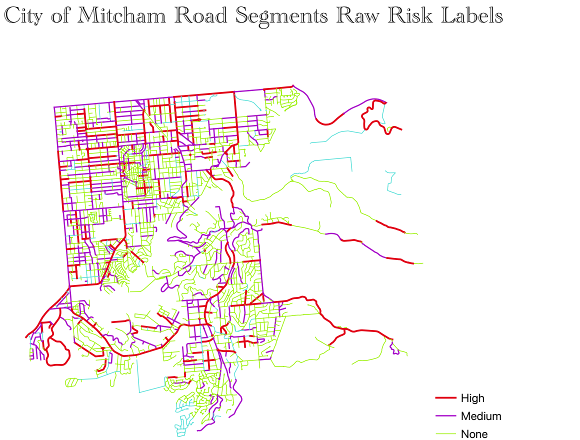


# Results: Risk classification

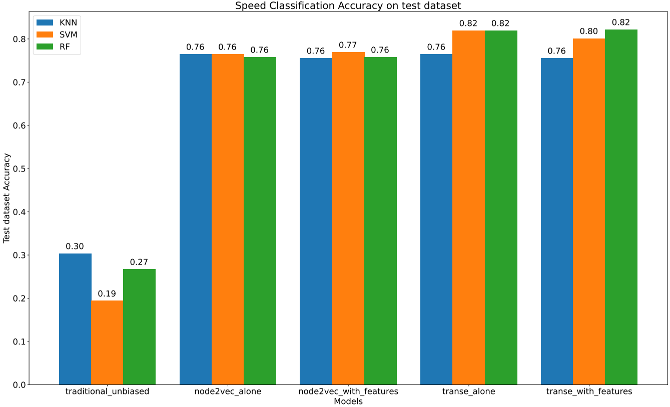



Map visualisation of road segments risk levels predicted by node2vec + kNN

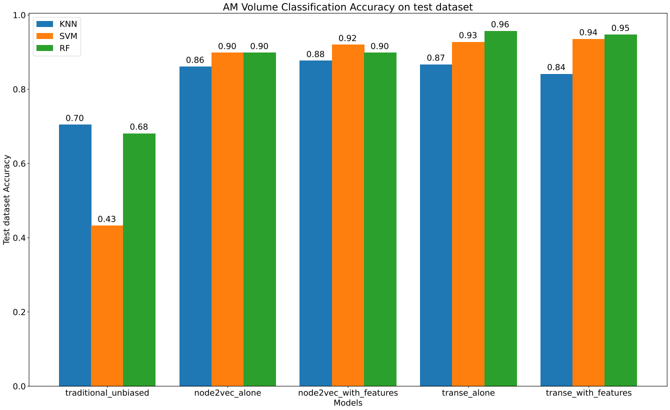
# Results: Risk classification



# Results: Speed classification



# Results: AM Volume classification



# Conclusion: Traffic

- Speed, volume and accidents are seem to be mainly determined by the road network structure.
- Road networks are not the whole story, with help of traffic knowledge graphs, classification accuracy can further improve.
- *node2vec_alone* and *transe_alone* models can handle missing data

# Conclusion: Spatial datasets

- Represent information with graphs or knowledge graphs, can encode structural information. Furthermore, it can handle missing data problem.
- Effective use of the spatial information in the dataset can effectively improve the accuracy of the classification tasks.
- Knowledge graph embeddings can preseve more information, so perform slightly better than homogenous network embeddings.
- Graph and Knowledge Graph embeddings perform well for Non-I.I.D Data feature extraction.

# Future work

- Traffic related
  - Apply on additional traffic related datasets to validate findings.
  - Build more complex traffic knowledge graphs, evaluate the performance.
- Spatial datasets related
  - Build more diverse spatial knowledge graphs.
  - Apply knowledge graph embedding models other than TransE.
- Other
  - Make use of the edge embeddings.
  - Explore how to represent and interpretate temporal information.

THANK YOU

Q & A