Forecasting Inter-destination Tourism Flow Within City With Social Media Data via Hybrid Learning model

--- benchmark, interpretation and model

Speaker: Hanxi Fang

Instructor: Prof. Song Gao

CONTENTS

PART ONE

Introduction

Introduction

https://www.mafengwo.cn/xc/10065/

When planning their trip how people decide which scenic spots to go in the same trip?

Specifically, why people will go to both scenic spot A and B in the same trip?

- Near?
- Of the same type?
- They are both popular?
- Drop by visiting?
- ……

Definition of “Inter-destination Tourism Flow”(ITF) :

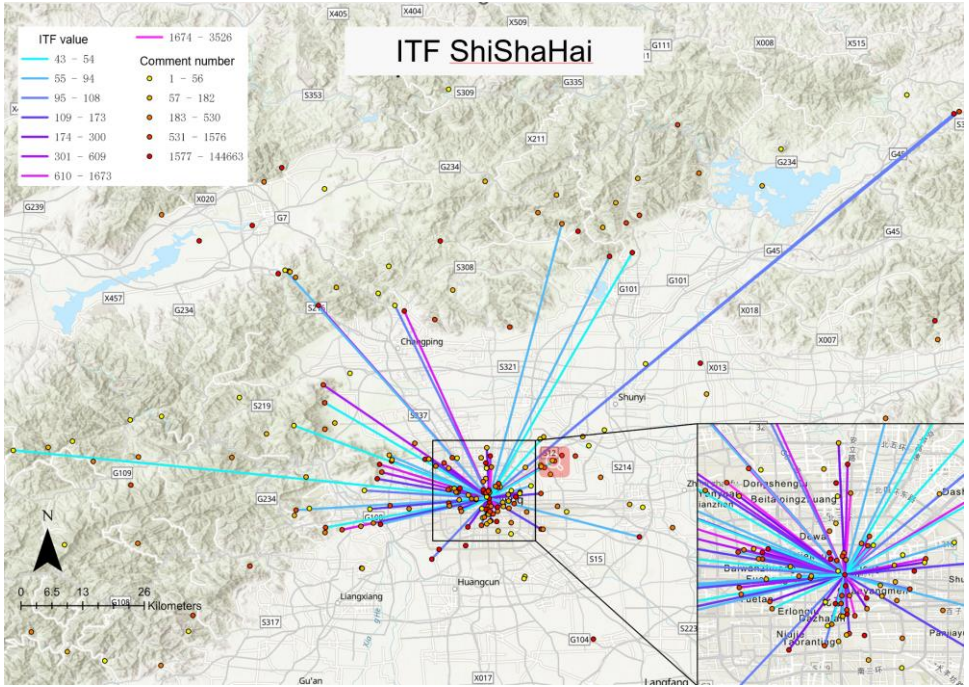
The ITF between scenic spot A and B refers to the total number of trips that visit both A and B.

The nature of ITF:
A quantitative metric to represent the flow intensity between tourism destinations.

In this research, we collect over 60000 trip diaries to Beijing from MaFengWo.com.

We extract over 20000 trips from them.

And On the basis of it, we get the ITF between any two scenic spots in the total 245 scenic spots in Beijing.



Usage of ITF

For tourism companies:

- Help to design a tourism recommending system and combined tickets for scenic spots.

for city planners:

- Offer important guide to the design of city tour bus.
- Offer guide to city planning, especially tourism planning that regard the whole city as a system.

for scholars:

- Define the interaction between different scenic spot, thus help us to understand the spatial pattern of tourism within city

Characteristics of ITF

ITF is hard to get

The formation of ITF is complex and hard to explain

Contribution

first

We put forward the concept of ITF, which quantitatively describe the interaction between different scenic spots within the same city; and create a benchmark about it from multiple sources of social media data

second

We create a hybrid GNN-based learning model that can predict the ITF . And the result has a mean absolute error less than 53%

third

We give an interpretation of the relationship between features of scenic spots and ITF.

The features include both explicit features of a single scenic spots, and the graph structure features of the interaction graph.

PART TWO

Relevant Works

Tourism Flow Research

01

02

03

The research scale are always the flow between cities or even countries.

Crampon, L.J. and Tan, K.T. (1973)

Put forward a regression model to predict international tourism flow between countries

Yang, X. A. , et al.(2020)

captures the nationality and movement patterns of foreign tourists to South Korea, and use a community detection algorithm partitions based on tourism flow between cities

Seok, H., Barnett, G.A. & Nam, Y(2021)

Use social network data to analyze international tourism flow.

… …

Flow Generation Problem

01 Gravity model, deep gravity model...

02 Random forest, XGBoost...

03 SIGCN, RFGCN...

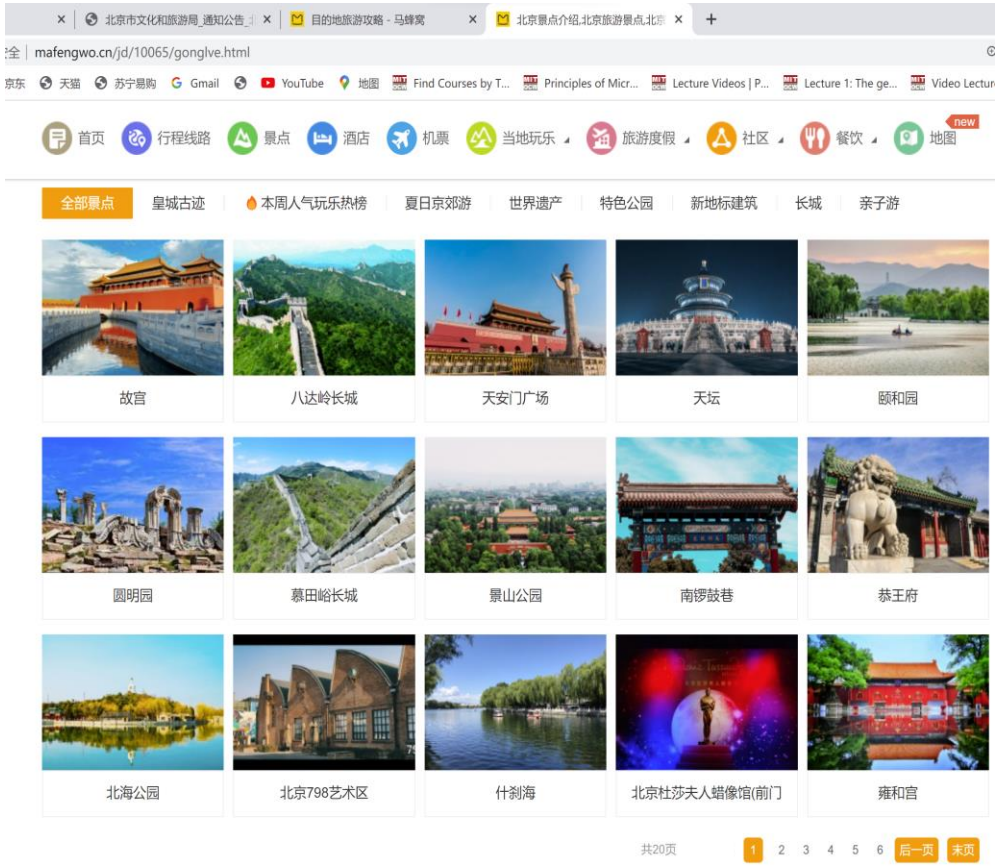
... ...

The researches are mainly focus on predicting residence-work commute flow, the data are usually less noisy with low variance.

PART THREE

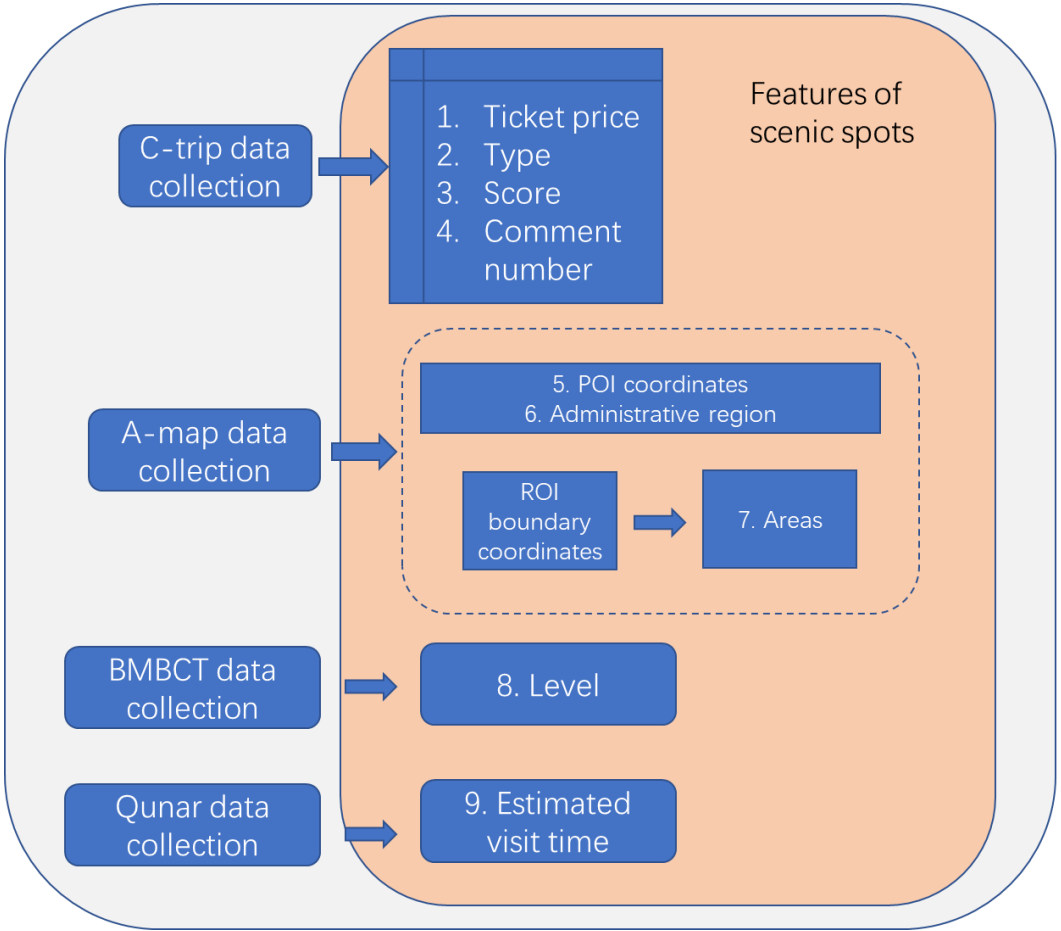
Data & Methods

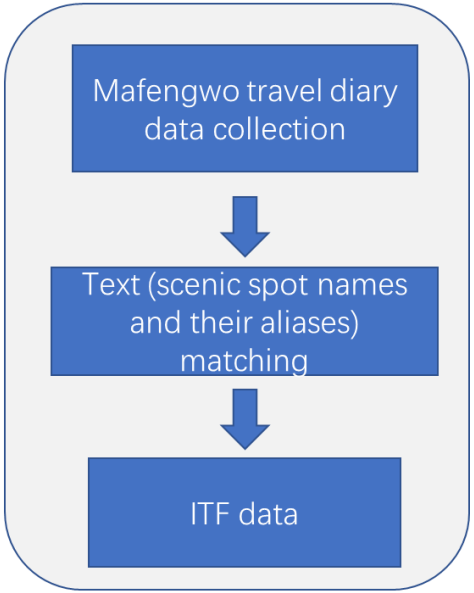
Data source

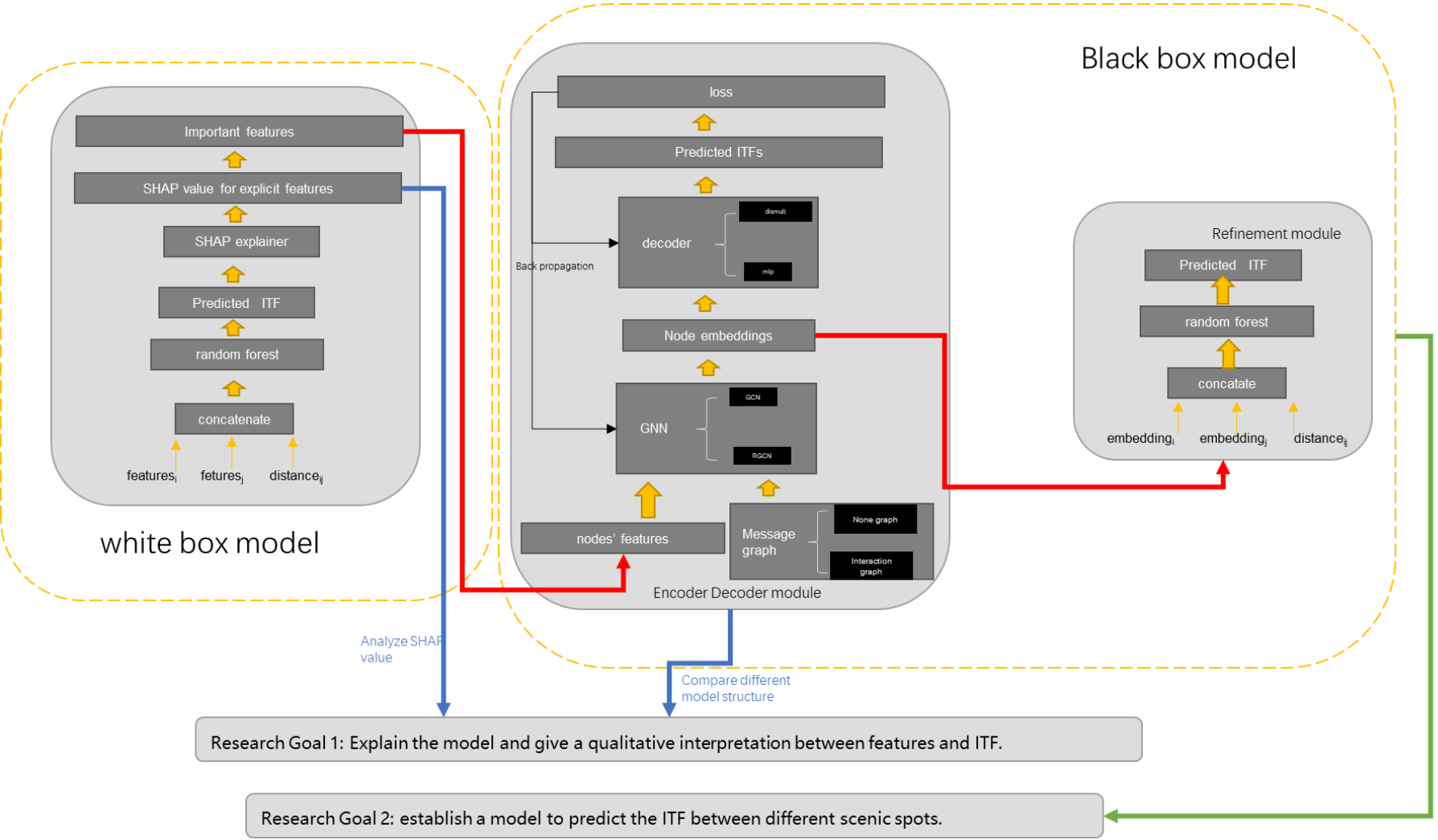


- MaFengWo: http://www.mafengwo.cn/
- Amap: https://www.amap.com/
- Ctrip: https://ctrip.com/
- Qunar: https://travel.qunar.com/
- Beijing Municipal Bureau of Culture and Tourism: http://whlyj.beijing.gov.cn/

Data Collection Procedure






Black box model
Important features
SHAP value for explicit features
SHAP explainer
Predicted ITF
random forest
concatenate
features$_i$
fetures$_j$
distance$_{ij}$
white box model
loss
Predicted ITFs
decoder
dismult
mlp
Back propagation
Node embeddings
GNN
GCN
RGCN
nodes' features
Message graph
None graph
Interaction graph
Encoder Decoder module
Refinement module
Predicted ITF
random forest
concatate
embedding$_i$
embedding$_j$
distance$_{ij}$
Analyze SHAP value
Compare different model structure
Research Goal 1: Explain the model and give a qualitative interpretation between features and ITF.
Research Goal 2: establish a model to predict the ITF between different scenic spots.

PART FOUR

Results & Conclusion

Random forest----Multicollinearity testing

The correlation between continuous variables: Use VIF to test, VIF less then 10 means no obvious Multicollinearity.

Feature	VIF
Score	4.48
Price	1.4
Comment number	1.31
Area	1.12
Level	4.12
Estimated time	4.24

VIF values

The correlation between catogorical valuables: Use Cramér's Vs to test, Cramér's Vs less then 5 means no obvious Multicollinearity.

	adname	Type
Adname	1.0	0.4
Type	0.4	1.0

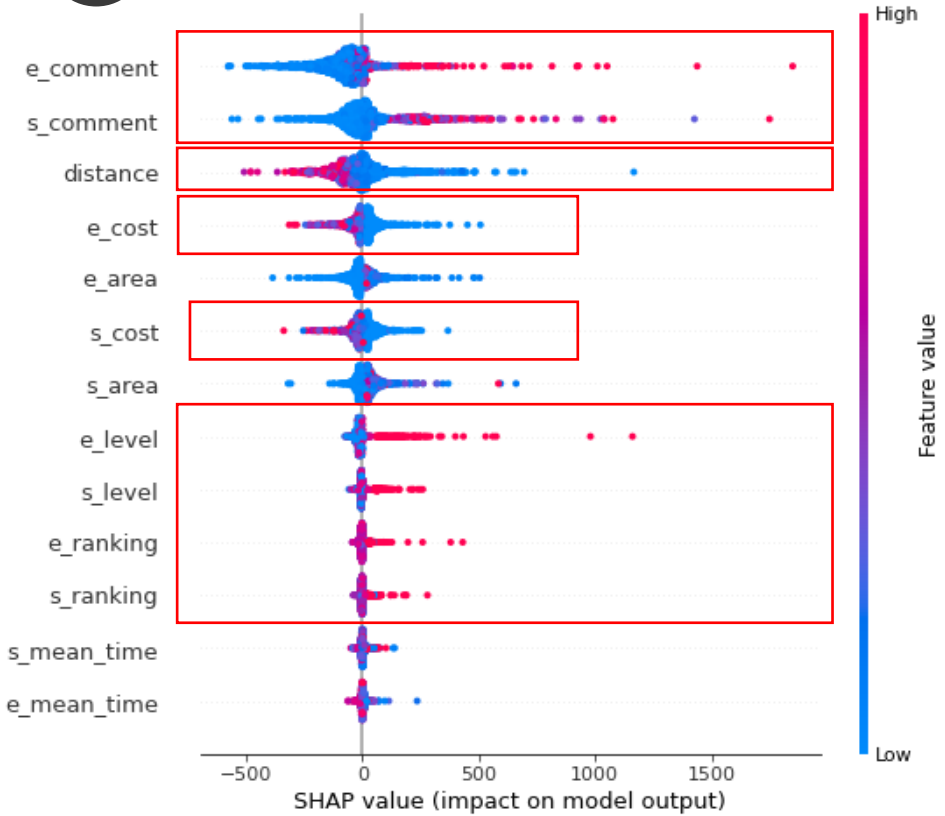
Cramér's Vs correlation value

The correlation between catogorical valuable and continuous valuables : Use One-way ANOVA test to get a p-value which means the probability of no correlation

	Score	Price	Comment number	Level	Area	Estimated time
Type	0.0	0.0	0.0	0.0	0.0	0.0
adname	1.62e-148	4.60e-288	5.60e-95	2.07e-194	0.0	0.0

P-value of one-way ANOVA test

Random forest + shap



Problem of random forest:

1. Hard to consider the corresponding features of start spot and end spot jointly.

2. It views the prediction of all the ITF as irrelevant regression problem, thus cannot take interaction graph structure of the spots into consideration.

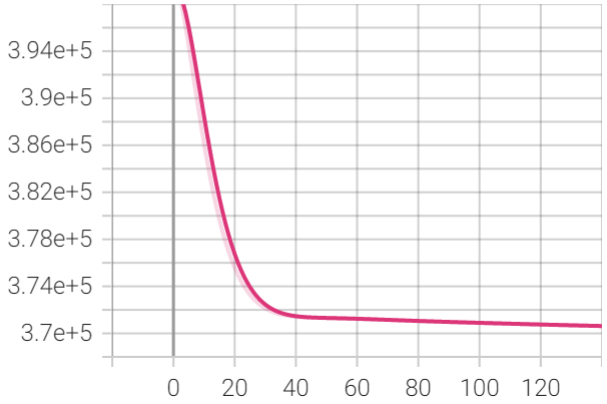
3. The method doesn't make sure that the ITF between A to B is equal to the ITF between B and A; which is true in reality.

Test mape: 1.915

S_: features of the start point; E_:features of the end point

Deep Gravity

train
tag: Loss/train



Practically, the deep gravity model is hard to train for this problem!

Problem of random forest:

1. solved
2. It views all the ITFs as irrelevant regression problem, thus cannot take spatial structure of the spots into consideration.
3. The method doesn't make sure that the ITF between A to B is equal to the ITF between B and A; which is true in reality.

BUT!!!Test mape:
5.7362

A Perfect Solution: GNN based models

GNN based model use the node features (the spot features) and the graph structure to generate an embedding for each node.

solve

After getting the embeddings of both node A and B, we can use DISMULT as the decoder to get the predicted ITF.

solve

solve

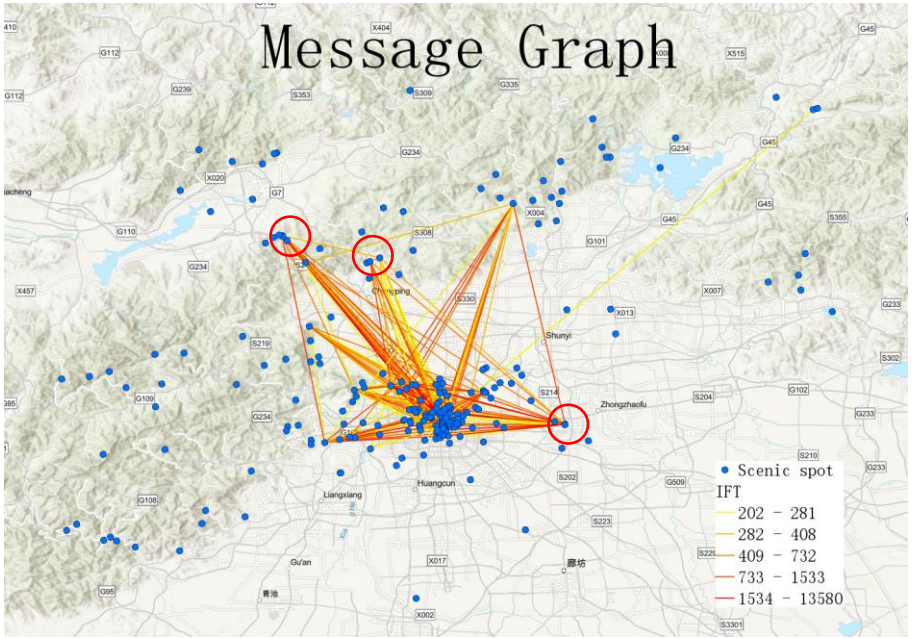
Dismult: a function (2 embeddings→ 1 scalar value)

Embedding of A: (1,2,3,4,5)
Embedding of B: (4,5,3,4,5)
Result = (1 * t1 * 4) + (2 * t2 * 5) + (3 * t3 * 6) + (4 * t4 * 4) + (5 * t5 * 5)
t1,t2,t3,t4,t5 are trainable parameters.

Problem of random forest:

1. It views all the ITFs as irrelevant regression problem, thus cannot take spatial structure of the spots into consideration.
2. Hard to consider the corresponding features of start spot and end spot jointly.
3. The method doesn't make sure that the ITF between A to B is equal to the ITF between B and A; which is true in reality.

GNN based models---message graph

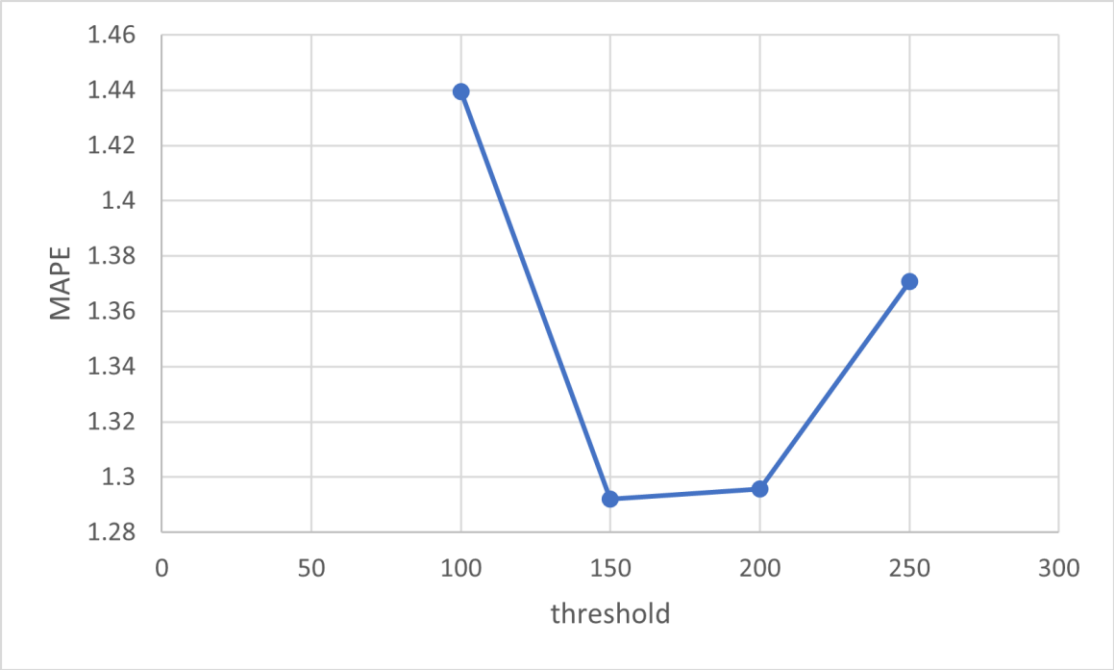


The message graph is a unweighted undirected graph. The existence of an edge between two spots A and B means there exists a strong interaction between A and B.

In our study, if there are more than 150 ITF between A and B in the training data, then there is an edge.

Note that when there is no initial information about any ITF between any two points, you can use other criterions to judge if there exists an edge, like commercial tour plans from tourism company; Since in fact the only thing you need to know is if there exist a large number of visitors who will visit both A and B, instead of the accurate number of the ITF.

GNN based models---threshold sensitive analysis



GNN based models---results

id	MODEL	USE_GRAPH	USE_DISMULT	MAPE	MSE
1	pure_rf	no	yes	1.9147	112085.5
2	deep_network(deep_gravity)	no	yes	5.7362	351221.8
3	rgcn+mlp	yes	no	5.7326	351222.1
4	gcn+mlp	yes	no	5.7321	351222.0
5	**rgcn+dismult(SIGCN)**	**yes**	**yes**	**1.3517**	**13711.59**
6	**gcn+dismult**	**yes**	**yes**	**1.7882**	**33560.63**
7	rgcn+with_only_self_edges	no	yes	2.8495	84514.1797

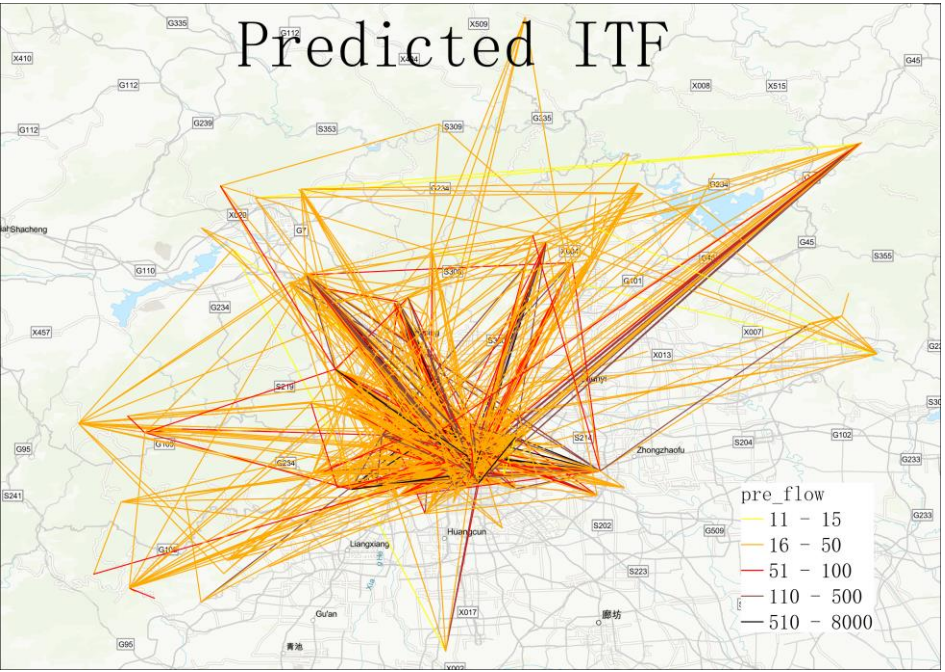
Conclution:

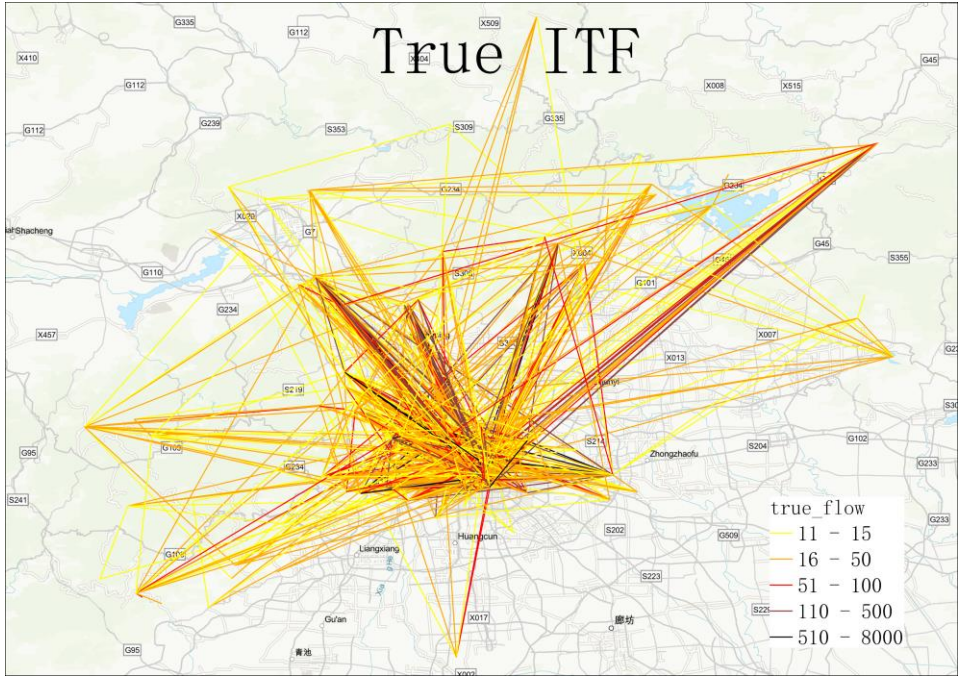
- 1/2 vs 5/6: gcn based model can be much better than RF or Deep_gravity.
- 3/4 vs 5/6: dismult makes sure symmetric result and manual correspondence of features, and is a key to the result.
- 7 vs 5: take the spatial interaction structure into account is also useful and important.

GCN+dismult+RF_refinement

ID	MODEL	USE RF as refininement	MSE	MAPE
1	rgcn+dismult(SIGCN)	no	13711.59	1.3517
2	gcn+dismult	no	33560.63	1.7882
3	**rgcn+dismult+RF**	**yes**	**21406.1004**	**0.549**
4	**gcn+dismult+RF**	**yes**	**17436.8982**	**0.5292**

GCN+dismult+RF_refinement





Conclusion

- The factors that influence the value of ITF between two spots can be divided by two aspects. The first aspect is the features of the two spots. According to the SHAP value of the random forest model, these mainly include the popularity(represented by comment number), ticket price, scores and level of both spots, and the distance between them. However, other features like tour spots' area and estimated visiting time have little impact.

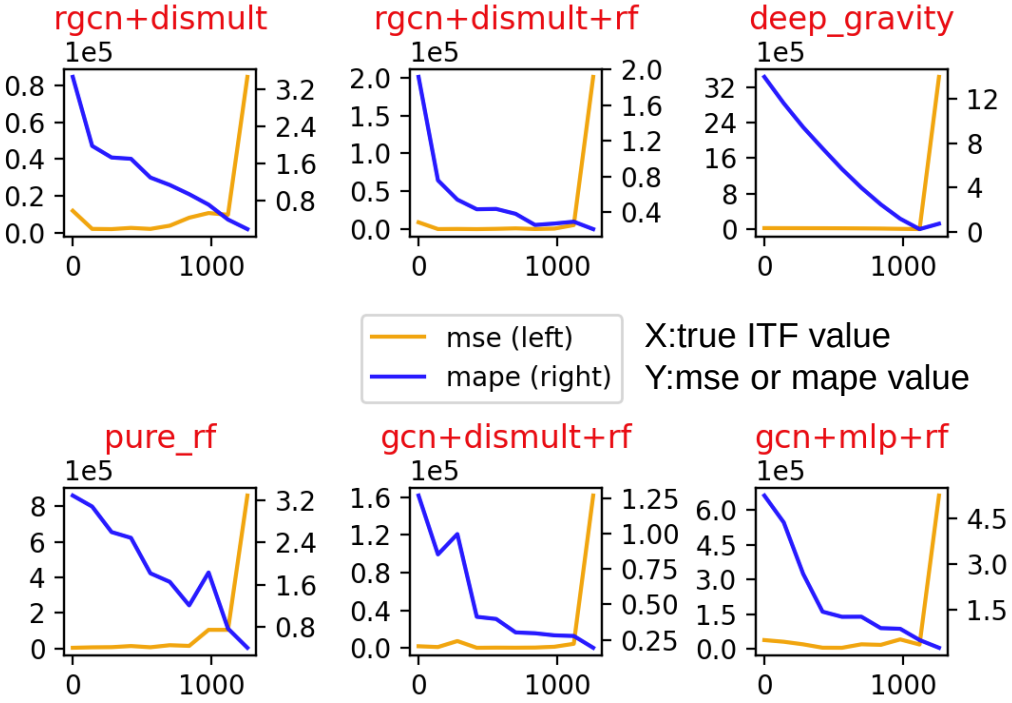
  The second aspect is the interaction graph structure influence the ITF. Taking account of the features of other scenic spots that have strong connection with the enquired spots by GNN can significantly improve the performance of the prediction model.

- For GNN based model, after getting the embedding for each spot, the usage of DISMULT decoder and the random forest as a refinement can significantly improve the performance of the model.

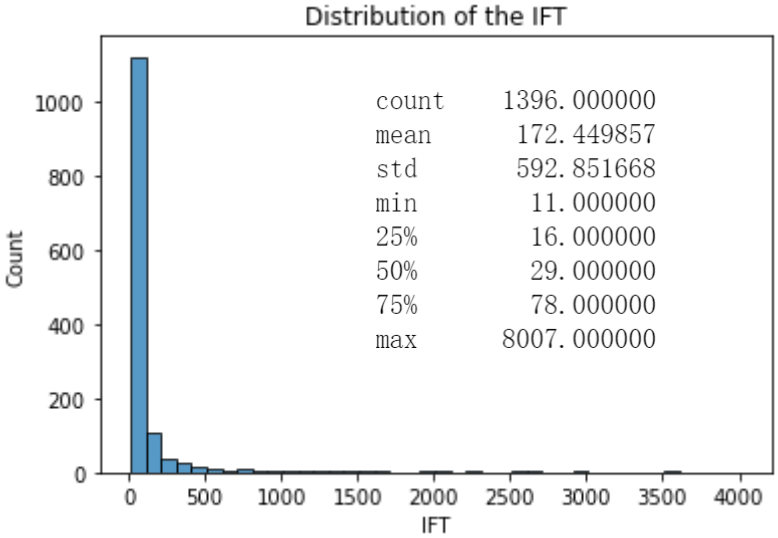
PART FIVE

Discussion

The drawbacks of the model



Due to the strong variance of ITF, different ITF value level has different precision.



Other Things to be discovered deeper

添加标题

1. How to get the information about where the user has gone from travel diary on social media with more precision.

2. How to explain how the structure of interaction graph influence the value of ITF.

3. More features can be used, like the feature embedding extracted from comment content data.

THANKS!