Causal Learning Empowered OD Prediction for Urban Planning

Jinwei Zeng
Tsinghua University
Beijing, China
zengjw17@gmail.com

Guozhen Zhang
Tsinghua University
Beijing, China
zgz18@mails.tsinghua.edu.cn

Can Rong
Tsinghua University
Beijing, China
rc20@mails.tsinghua.edu.cn

Jingtao Ding
Tsinghua University
Beijing, China
dingjt15@tsinghua.org.cn

Jian Yuan
Tsinghua University
Beijing, China
jyuan@tsinghua.edu.cn

Yong Li
Tsinghua University
Beijing, China
liyong07@tsinghua.edu.cn

ABSTRACT

Predicting future origin-destination (OD) flow is essential for urban planning since it provides feedback for planning adjustment and reference for road planning. However, OD prediction for urban planning scenarios is unique as it typically lacks training data. A common practice is to refer to data from other cities, which causes the out-of-distribution (OOD) problem. A promising solution is to leverage causal information in the data. However, there are two challenges in utilizing causal information in urban planning scenarios: (a) Urban system has numerous factors, and only part of them indicate causal information. (b) The planned city development correlates with original city characteristics, therefore bringing confounding bias to the causal modelling process. In this paper, we propose designs to solve both challenges. Specifically, we first design a causal disentangled representation module to identify causal factors in attributes. Second, we adopt a variational sample re-weighting module to reduce the confounding bias. Our proposed model outperforms seven state-of-the-art baselines on three real-world datasets, achieving an average improvement of 9.59% in the MAE metric. Further in-depth analysis shows our method's robustness across different urban planning scenarios and outstanding performance in predicting extremely large OD flows, which corroborates the contribution of our designs to the urban planning field.

CCS CONCEPTS

• **Applied computing** → **Forecasting**; • **Computing methodologies** → *Causal reasoning and diagnostics.*

KEYWORDS

causal learning, urban planning, OD prediction

ACM Reference Format:
Jinwei Zeng, Guozhen Zhang, Can Rong, Jingtao Ding, Jian Yuan, and Yong Li. 2022. Causal Learning Empowered OD Prediction for Urban Planning.

CIKM '22, October 17–21, 2022, Atlanta, GA, USA.

ACM ISBN 978-1-4503-9236-5/22/10…$15.00
https://doi.org/10.1145/3511808.3557255

In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3511808.3557255

1 INTRODUCTION

Cities are still expanding and renewing rapidly nowadays [7]. To ensure the rational and efficient use of land, urban planning has received increasing research attention [26]. Predicting future origin-destination flow under planned city development is an essential problem in urban planning. For one aspect, people's movement in the city indicates how their demand for facilities is met. Therefore planners can refer to the city accessibility exhibited by predicted future city OD to adjust urban planning policy. For another, the transportation planning domain also relies on predicted future OD to plan future city transportation in advance [23]. Therefore, the precise prediction of future OD flow is valuable for multiple aspects of urban planning.

There have been many solutions to the general OD prediction problem. These attempts can mainly be divided into model-based methods and data-driven methods. Model-based methods typically attribute people's physical mobility patterns to some intrinsic mechanism, such as the gravity model [39] and the intervening opportunity model [31, 32]. Such model-based models have good explainability but show limited performance for only including basic region features. With the fast development of machine learning and deep learning, data-driven methods are introduced which utilize the correlation between features and target [22, 30] to make predictions. Liu et al. [19, 37] further introduce graph methods that characterize the adjacency of regions to learn region representation.

However, the OD prediction task for urban planning scenarios differs from the traditional one. Since, in urban planning scenarios, planners focus on how planned city development will influence OD, the OD prediction can therefore be regarded as a counterfactual regression where data-driven methods will bring confounding bias. Moreover, in practical applications, planners suffer from the absence of historical data since urban planning is primarily required in less developed regions where the data reserve capacity is weak. Therefore, planners usually rely on development data from other regions as a reference. This way, the attribute distribution distinctions between regions introduce the out-of-distribution (OOD) problem. Since existing data-driven trials are all correlation-based, it is hard for these methods to adapt to such an OOD situation. Therefore, nowadays, the urban planning domain still depends on traditional

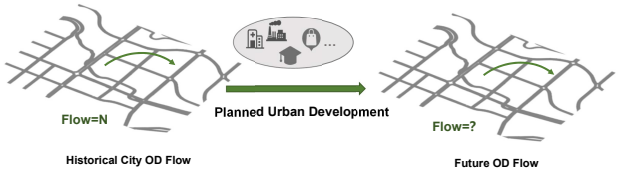

Figure 1: OD prediction for urban planning, which is to predict future city OD based on historical city OD and the planned city development.

model-based methods to predict the future OD [2, 21, 23], resulting in limited prediction performance.

Considering the specialities of OD prediction for urban planning, we turn to causal learning prediction methods for solutions. Firstly, causal learning methods are better at learning causal information, thereby predicting the counterfactual outcome more accurately. Secondly, causal methods can capture true causal relationships and better generalize to the OOD situations. Since causal relationships are mostly robust across the training set and testing set even in an OOD situation, stable causal relationships rather than spurious correlation inside the dataset help the model tackle the OOD problem [8]. To conclude, it is promising to introduce causal learning methods to our task. However, there are three challenges for us in incorporating causal methods: (1) The treatment, which is the planned city development in our task, is a bundle continuous treatment and hard to model. Since we want to model how planned development will impact future city OD, we have to model the causal influence of planned development, which is the treatment in causal terms. While existing causal prediction methods only deal with binary treatment cases, the planned development is usually depicted by the change in the count of all kinds of POIs, a bundle continuous treatment. Modeling bundle continuous treatment is challenging since we cannot go the way of defining treatment groups and balancing between different groups [40]. (2) Our treatment is heavily confounded by region attributes because the treatment, which is the planned city development in our task, is not randomly assigned to every region but largely depends on region attributes [1]. (3) The urban system has various factors [5], but not all these factors causally affect city OD. Some of the factors may even cause bias in the prediction of the OOD situation. Therefore, it is essential to find the underlying causal factors and predict OD flows only on them.

To deal with these challenges, we propose an OD prediction model with cauSal dIsentangled REpresentation and re-weIghting (SIRI)[1]. Specifically, we first design the causal disentangled representation module to identify the causal factors of region attributes in an adversarial learning way. The underlying factors can be dissected into instrumental, confounding, and adjusting factors. We only predict our target by confounding and adjusting variables since they causally affect the target. Secondly, we introduce the variational sample re-weighting module to ease the confounding bias brought by confounding factors. Finally, we combine these two modules in an end-to-end way and conduct extensive experiments to validate our method's effectiveness. Further in-depth analysis shows our method's consistent superiority in various urban planning scenarios and outstanding performance in predicting extremely large OD flows, which corroborate our work's value to urban planning.

We highlight our contributions as follows:

- We propose a novel OD prediction method for practical use in urban planning scenarios, which can predict future OD distribution influenced by planned city development. Since OD prediction for the urban planning domain is an important task but faces unique challenges, our work successfully leverages causal information to tackle the challenges.
- We design an end-to-end solution with two novel causal modules, the causal disentangled representation module and the variational re-weighting module, to learn the causal information and leverage it to predict future OD impacted by the planned city development. Our model's capability of handling bundle continuous treatment, which is the planned city development in our task, is also remarkable progress for the causal learning field.
- We conduct extensive experiments on three real-world datasets from different cities and with different scales. Experimental results show that our model outperforms existing state-of-the-art methods by 9.59% on average in the MAE metric. Further in-depth analysis shows our method's consistent superiority over all urban planning scenarios and extraordinary ability to predict extremely large OD flows.

2 PROBLEM DEFINITION

Our goal is to design a causally empowered OD prediction framework for urban planning scenarios as shown in Figure 1. We regard POI changes in count as the expression form of planned city development and want to predict future city OD under the planned POI changes. Specifically, we are given historical OD flow matrix X_f, region attributes X_a and POI changes T corresponding to the urban planning policies, and our task is to predict future OD flow matrix Y. The task can be formulated as below,

$$Y = F(X_f, X_a, T), \tag{1}$$

where $X_f, Y \in \mathbb{R}^{n\times n}$, $X_a \in \mathbb{R}^{n\times n_a}$, $T \in \mathbb{R}^{n\times n_t}$, with n, n_a and n_t denoting number of regions, dim of region attributes and dim of POI changes, respectively.

Since we want to model how POI changes impact future city OD distribution, POI changes T are the treatment variables in our task in the context of causal learning [20].

3 METHODOLOGY

3.1 OD Prediction Backbone Model

Since our problem is to predict the future city OD flow under planned city development, we adopt a well-acknowledged OD prediction framework, DeepGravity [30], as our backbone model on which we make causal adaptations to fit urban planning scenarios.

Design of the backbone model is shown in Figure 2(a). This model interprets OD prediction as an outflow allocation task and predicts the destination probabilities of flows departing from an origin. As we can see in the graph, for every OD pair, the features X and the POI changes T of both the origin and the destination are fed into a feedforward network to output a score. By applying softmax on the

[1] https://github.com/tsinghua-fib-lab/SIRI

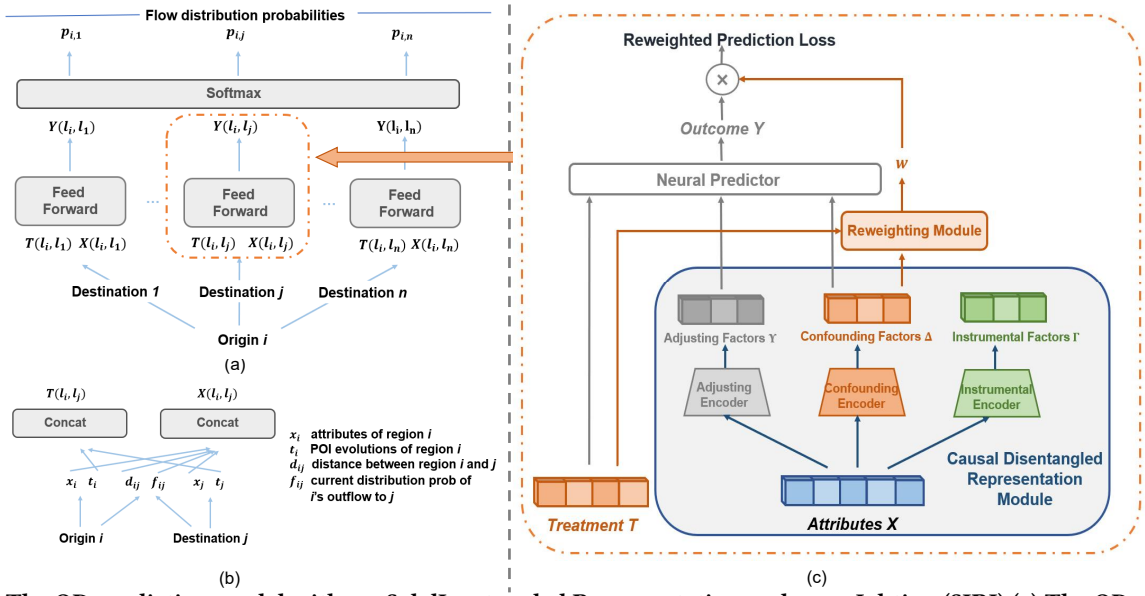

Figure 2: The OD prediction model with cauSal dIsentangled Representation and re-weIghting (SIRI) (a) The OD prediction backbone model: DeepGravity. (b) Inputs for the feedforward network. (c) Our proposed causal feedforward network, SIRI.

scores of all OD pairs originating at O, we get the probabilities of O's outflow going to respective destinations. Finally, the OD flow is available by multiplying the outflow of origin i by the probability of it ending at destination j.

The inputs of the feedforward network in our task are shown in Figure 2(b). For every OD pair (i, j) fed into this network, we input the attributes of origin i and destination j, the distance between origin and destination d_{ij} and the historical flow of this OD pair f_{ij} as environmental variables $X(l_i, l_j)$. As both POI changes of the origin and the destination will impact the OD flow, we aggregate the planned POI changes of the origin and destination as the treatment variables $T(l_i, l_j)$. $X(l_i, l_j)$ and $T(l_i, l_j)$ constitute the overall inputs of our feedforward network.

Based on this OD prediction backbone model, we make causal adjustments to the feedforward network, substituting it with our proposed SIRI model as shown in Figure 2(c). The SIRI model comprises two causal designs, the causal disentangled representation module and the variational re-weighting module. In the latter sections, we will elaborate on the SIRI model comprehensively.

3.2 Causal Disentangled Representation Module

The urban system can be regarded as a complex system with numerous factors and intricate relationship [14]. Although we can use all those numerous attributes to fully leverage all data correlation, not all these correlation is stable across training set and testing set. Since in urban planning scenarios, usually, regions to be planned are less-developed regions that barely have a good data reserve, planners have to train the model on developed regions with known development data. Therefore, the distribution gap between training data and testing data is further broadened, worsening the out-of-distribution(OOD) situation. Considering that causal relationships are generally salient and stable, basing the prediction on causal relationships saves the model from noise and bias brought by spurious correlation and helps the model better generalize to OOD

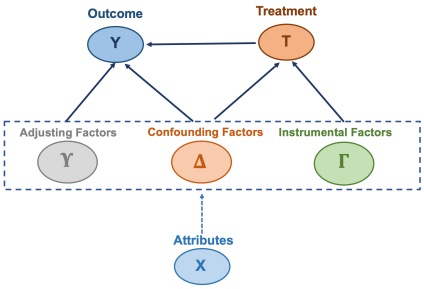

Figure 3: Causal graph of underlying factors of attributes to treatment and outcome.

situations [28]. In this sense, we design the causal disentangled presentation module to identify the underlying factors of attributes that causally affect the outcome and predict the outcome of these causal factors.

Without loss of generality, there are three types of underlying factors for attributes based on factors' causal relation to treatment and outcome [13]. We plot the causal graph of these factors with treatment and outcome in Figure 3. In causal terms, factors that only causally affect treatment are referred to as instrumental factors Γ, whereas factors that only causally affect the outcome are referred to as adjusting factors Υ. The factors that causally affect treatment and outcome simultaneously are referred to as confounding factors Δ. As we can see, only confounding factors and adjusting factors have a causal effect on the outcome. Therefore, our final prediction of the outcome should not include the instrumental factors.

In previous work, Hassanpour [13] proposed a disentangled representation module to identify the three types of underlying factors. However, this work is limited in handling binary treatment. Since the treatment in our case is bundle continuous treatment, we make several adaptations to this module to fit our scenario. Here we elaborate on our disentangled representation module. The sketch of the module is shown in Figure 4. The representation part

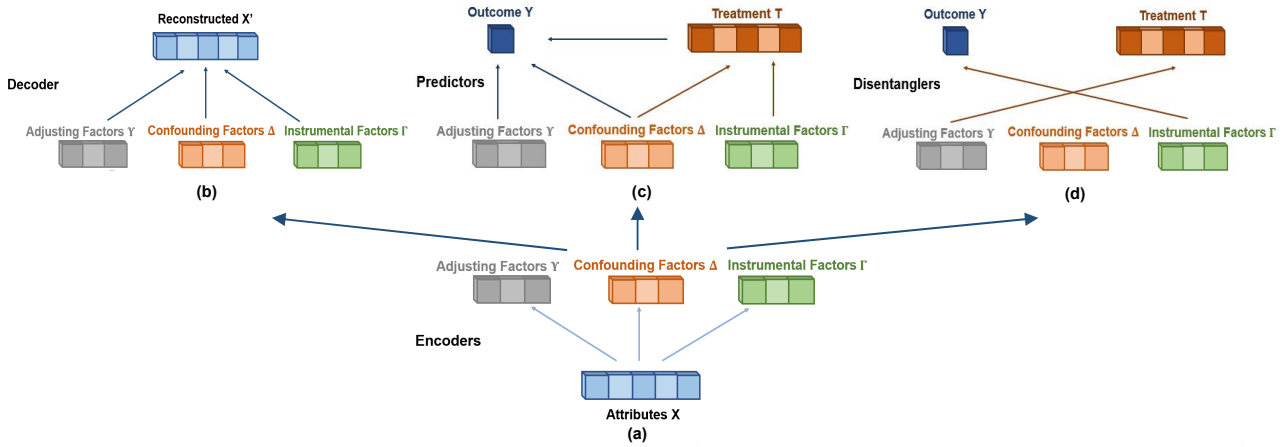

Figure 4: The causal disentangled representation module in SIRI. (a) The representation layer with three encoders. (b) The reconstruction decoder. (c) The causal regression predictors. (d) The causal disentangling predictors.

is shown in Figure 4(a). To ensure the correct identification of the representations, we propose downstream designs to regulate the optimization of representation part as shown in 4(b)(c)(d). Overall, our disentangled representation module is composed of:

Representation learning encoders as shown in all three subfigures in Figure 4: Each of them encodes one type of underlying factors, the adjusting factors $\Upsilon(x)$, the confounding factors $\Delta(x)$, and the instrumental factors $\Gamma(x)$.

Reconstruction decoder as shown in Figure 4(b): It reconstructs the input attributes X by encoded three underlying representations $\Gamma(x)$, $\Delta(x)$, $\Upsilon(x)$. The reconstruction of X is denoted as $H(\Gamma(x), \Delta(x), \Upsilon(x))$. This decoder is included to ensure that the three encoders grasp all the underlying factors of the attributes.

Regression networks 'predictors' as shown in Figure 4(c); The first predictor $H_1(\Gamma(x), \Delta(x))$ is to model the treatment. The second predictor $H_2(\Delta(x), \Upsilon(x), T)$ is to predict the outcome. Since $\Gamma(x)$ and $\Delta(x)$ are the only factors that causally affect T, $\Gamma(x)$ and $\Delta(x)$ should embed all information in attributes for treatment assignment and show a good performance in predicting treatment T. So as $\Delta(x)$, $\Upsilon(x)$ and T for outcome.

Although in this way, we can embed all factors causally affect T in $\Gamma(x), \Delta(x)$, and all factors causally affect Y in $\Delta(x), \Upsilon(x)$ and T, we have no guarantee that $\Gamma(x)$ only affects T and $\Upsilon(x)$ only affects Y. In other words, there must be some regularization on $\Upsilon(x)$ to be disentangled with outcome T, so as $\Gamma(x)$ to Y. We first work on disentangling $\Upsilon(x)$ and T. Existing disentangle methods are all limited in handling binary treatment, which typically adopt some distance metric to balance the distribution $Pr(Y|T=0)$ with $Pr(Y|T=1)$. However, there are too many conditional distributions $Pr(Y|T)$ when treatment is non-binary and continuous since T can take numerous values. Thereby, it is hard to apply existing distance metrics to disentangle bundle continuous treatment T with $\Upsilon(x)$.

Our solution to this dilemma is based on an assumption that if even the best predictor cannot predict T by $\Upsilon(x)$ well, then we can regard T and $\Upsilon(x)$ as disentangled. Therefore, we add components for disentangling to this module:

Regression networks 'disentanglers' as shown in Figure 4(d): $H3(\Gamma(x))$ is to predict the outcome Y by $\Gamma(x)$. $H4(\Upsilon(x))$ is to predict the treatment T by $\Upsilon(x)$. Guided by the assumption, we achieve the disentanglement in an adversarial way. That is, take disentangling T and $\Upsilon(x)$ for illustration. At every iteration of training, we first maximize the loss of regressing T by $\Upsilon(x)$ so that the underlying factors $\Upsilon(x)$ will adjust to the representation space that disentangles with T. Then we fit the regressor to this new representation space by minimizing the loss of regressing T by $\Upsilon(x)$. When the regressor fails to predict T by $\Upsilon(x)$ however we minimize the prediction loss, we think $\Upsilon(x)$ embeds no information about T and they are disentangled. We can apply the same procedure to disentangle Γ with Y. The loss terms in this module are formulated as below:

$$
\begin{aligned}
L_r &= \frac{1}{N}\sum_{i=1}^{N} L(x_i, H(\Gamma(x_i), \Delta(x_i), \Upsilon(x_i))) \\
L_{p,1} &= \frac{1}{N}\sum_{i=1}^{N} L(y_i, H_2(\Delta(x_i), \Upsilon(x_i))) \\
L_{p,2} &= \frac{1}{N}\sum_{i=1}^{N} L_(t_i, H_1(\Gamma(x_i), \Delta(x_i))) \\
L_{d,1} &= \frac{1}{N}\sum_{i=1}^{N} L_(y_i, H3(\Gamma(x_i))) \\
L_{d,2} &= \frac{1}{N}\sum_{i=1}^{N} L_(t_i, H4(\Upsilon(x_i))).
\end{aligned} \tag{2}
$$

Therefore, the overall training objective of our causal disentangled representation module can be formulated as iteratively minimizing these two adversarial losses:

$$
L(\Gamma, \Delta, \Upsilon, H, H_1, H_2) = L_r + \alpha \cdot L_{p,1} + \beta \cdot L_{p,2} - \gamma \cdot L_{d,1} - \delta \cdot L_{d,2} \tag{3}
$$

$$
L(H_3, H_4) = \gamma \cdot L_{d,1} + \delta \cdot L_{d,2} \tag{4}
$$

where α, β, γ and δ are the hyper-parameters to balance these losses.

In this way, we disentangle our attributes to three causal types of underlying factors. Not only can we base prediction on the real causal factors, $\Delta$and Υ to get better generalizability, but we identify the exact confounding factors Δ, which makes for the latter deconfounding part.

3.3 Variational Re-weighting Module for Deconfounding

Since our task is to predict future city OD under planned city development, we want to know what the OD distribution will be like when the city POIs are distributed in a new way. In this sense, our problem answers a counterfactual question since we can only clearly infer the impacted OD flows when we have precise command of how POI changes affect OD flows. However, counterfactual regression usually suffers from treatment confoundedness. Generally, the treatment, which is the planned POI changes, is never randomly assigned: urban planning policies depends on the local city development and conditions [1]. For example, we will design more amenities for an ageing city, but this case may not necessarily generalize to other cities. In this way, the change of amenity POIs is confounded by citizen age where we cannot differentiate the influence of amenity POI changes and citizen average age on OD flows. Therefore, we can conclude that making counterfactual prediction by observational data induces confounding bias [40], resulting in the model failing to infer impacted future OD flow.

An acknowledged solution is to decorrelate attributes and treatments in the observational dataset. While Randomized Controlled Trials [11] is the most intuitive method for causal inference, such randomization is unrealizable in real practice. Previous work generally adopts re-weighting methods [3, 25] to put larger weights on instances with treatment that is rarely assigned and smaller weights on instances with common treatment. The propensity score is incorporated to measure the treatment assignment probabilities. However, almost all these re-weighting methods are limited in dealing with binary treatment, except that generalized propensity score-based re-weighting [9] extends the solution to multi-valued treatment cases, and Zou [40] proposed a variational sample re-weighting method to deal with bundle binary treatment. But our case is even more complex since we have to handle bundle continuous treatment. An urban planning policy not only includes the adjustment of various types of POIs, but for every type of POI, the adjustment is typically a numerical value, referring to the change of POI numbers. However, we find that the mathematical deduction in [40] can be extended to the cases where bundle treatments are continuous. Therefore, we introduce this method to our case to realize deconfounding.

As shown in Figure 5, this method first adopts a variational autoencoder to learn the low dimensional latent representation Z of T. The distribution $p(Z)$ of the latent treatment representation denotes the prior distribution of the latent treatment representation and is usually assumed to follow standard normal distribution $N(0, I)$. After getting the distribution of Z, we can deduce the re-weighting weights as below,

$$
\begin{aligned}
{w_i}^d =& \frac{p(t_i)}{p(t_i|x_i)} = \frac{p(t_i)}{\int_z p(z|x_i)p(t_i|z)dz} = \frac{1}{\int_z p(z|x_i)\frac{p(t_i|z)}{p(t_i)}dz} \\
=& \frac{1}{\int_z p(z|x_i)\frac{p(z|t_i)}{p(z)}dz} = \frac{1}{\int_z p(z|t_i)\frac{p(z|x_i)}{p(z)}dz} \\
=& \frac{1}{\int_z p(z|t_i)\frac{p(z,x_i)}{p(z)p(x_i)}dz}.
\end{aligned}
$$

To calculate $\frac{p(z,x_i)}{p(z)p(x_i)}$, this method introduces a neural network based classifier $p_{L|X,Z}$ to estimate how much z and x_i are entangled. We denote the entangled case as $L = 1$ and the disentangled case as $L = 0$. Then, applying Bayes' Rule [34], this method transform $\frac{p(z,x_i)}{p(z)p(x_i)}$ to a new form,

$$
\frac{p(z, x_i)}{p(z)p(x_i)} = \frac{p(z, x_i|L = 1)}{p(z, x_i|L = 0)} = \frac{p(L = 0)}{p(L = 1)} \frac{p(L = 1|z, x_i)}{p(L = 0|z, x_i)}. \tag{5}
$$

In this way, we transform the weights' calculation to classify $p(L|X, Z)$. As shown in Figure 5(a), we feed (X, Z) with labels of L into the classifier to train it, and used the trained classifier to get $p(L|X, Z)$ and to calculate $\frac{p(z,x_i)}{p(z)p(x_i)}$. Finally we take the results back to Equation 5 to calculate the final re-weighting sample weights.

After obtaining the sample weights, we add them to the prediction loss term $\frac{1}{N}\sum_{i=1}^{N} L_1(y_i, H_2(\Delta(x_i), \Upsilon(x_i)))$. The revised loss term $L(\Gamma, \Delta, \Upsilon, H_1, H_2)$ is as below,

$$
L_{p,1*} = \frac{1}{N}\sum_{i=1}^{N} {w_i}^d \cdot L(y_i, H_2(\Delta(x_i), \Upsilon(x_i))) \tag{6}
$$

$$
L(\Gamma, \Delta, \Upsilon, H, H_1, H_2) = L_r + \alpha \cdot L_{p,1*} + \beta \cdot L_{p,2} - \gamma \cdot L_{d,1} - \delta \cdot L_{d,2} \tag{7}
$$

In this way, we carefully decorrelate our treatment with the confounding factors, which helps our final OD flow predictor capture the causal effect of treatment and factors more precisely. Moreover, the causal constructing representation module enhances the deconfouding performance since it helps us identify the exact confounding factors underlying all attributes. Thereby, we can only decorrelate confounding factors with treatment without intervening in the prior distribution of treatment induced by instrumental factors.

3.4 End-to-end Training Framework

To fully tackle the challenges, we design the causal disentangled representation module to identify underlying causal factors of attributes and introduce the variational re-weighting module for deconfounding. Since the variational re-weighting module relies on the learned disentangled representation to calculate the weights more precisely, we split the training into two stages. During the first stage of the training, we train the OD prediction backbone model with the causal disentangled representation module. The training target is to iteratively minimize Equation (3) and Equation (4). After identifying the causal factors, we calculate the re-weighting weights based on Equation (5) and add the weights to the prediction loss. During the second stage, we continue training to minimize the re-weighted loss Equation (7) until convergence. In this way, we get an end-to-end training procedure as shown in Algorithm 1.

4 PERFORMANCE EVALUATION

We conduct extensive experiments on datasets from different cities and of different scales to fully evaluate our proposed method. We compare our method with seven existing state-of-the-art OD prediction methods. We also do the ablation study and in-depth analysis of our proposed method to validate our method further. In this section,

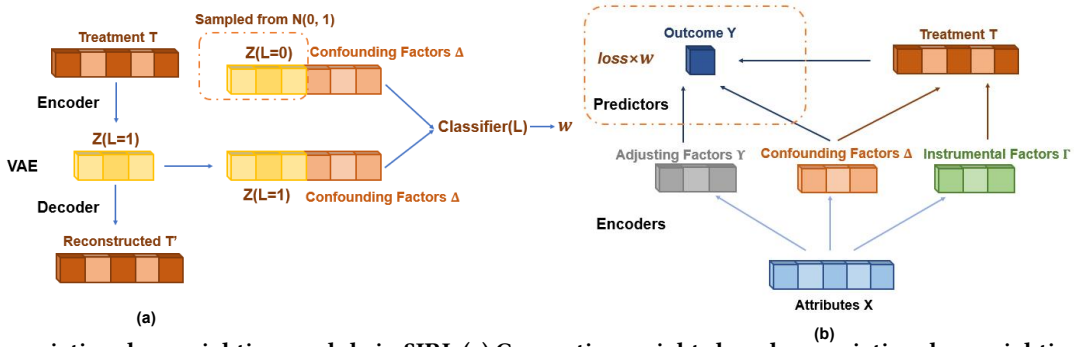

Figure 5: The variational re-weighting module in SIRI. (a) Computing weights based on variational re-weighting method. (b) Adding sample re-weighting weights to the outcome prediction loss in the disentangled representation module.

Algorithm 1 A causally empowered OD prediction model

Initialize encoders Γ, Δ, Υ, decoder H, predictors H_1, H_2, disentanglers H_3, H_4 in the feedforward network of the DeepGravity OD prediction framework.
// Learn the causal disentangled representations.
repeat
  Optimize loss Equation (3) and update $\Gamma, \Delta, \Upsilon, H, H_1, H_2$
  Optimize loss Equation (4) and update H_3, H_4
until loss Equation (3) < ϵ_1
// Learn the deconfounding sample weights.
Calculate sample weights w based on Equation (5) and (5).
// Add sample weights to the causal disentangled module and continue training.
repeat
  Optimize loss Equation (7) and update $\Gamma, \Delta, \Upsilon, H, H_1, H_2$
  Optimize loss Equation (4) and update H_3, H_4
until loss Equation (7) < ϵ_2

we first introduce our experimental setups, including the datasets, baseline methods, and evaluation protocols. Then, we elaborate on the experiment results and performance analysis.

4.1 Experiment Setups

4.1.1 Dataset. We conduct experiments on three real-world, large-scale datasets from different cities to validate our method. Reproducing practical urban planning scenarios, we train on developed regions with known historical data and predict on less-developed regions to be planned. The brief statistics of the three datasets can be referred to in Table 1.

Changsha Partial Dataset We collect data from Tianxin District and Yuhua District, the most developed areas in Changsha, as the training samples and predict on Wangcheng District and Changsha County. The latter two districts are developing at high speed these days. We have the attributes of these districts, the POI changes from Nov. 2019 to Nov. 2020, and the OD flow distribution in Nov. 2019 as input. Our target is to predict the OD flow distribution in Nov. 2020.

Changsha Global Dataset This dataset covers all urban districts in Changsha. The training set covers the five relatively more developed regions, Yuhua District, Kaifu District, Tianxin District, Yuelu District, and Furong District. The testing set includes the two relatively less developed regions, Wangcheng District and Changsha County. The training settings are the same as in the Changsha Partial Dataset.

Beijing Partial Dataset This dataset covers two large districts in Beijing, the Chaoyang district and the Tongzhou district, and the latter district is a recent planning and development emphasis in Beijing. In this dataset, we train on data from Chaoyang District and predict on Tongzhou District. We predict the OD flow distribution in Jul. 2021, given the OD flow distribution of Jul. 2020 and POI changes in this interval.

4.1.2 Baselines. To evaluate the performance of our model, we compare our framework with seven well-established methods from three groups.

Traditional Model-based Methods in the Urban Planning Field:

- Radiation Model [31]: This model assumes that mobility is determined by the population density distribution of the city. The relative probability of mobility from origin i to destination j is formulated as
$$\pi_{ij} = \frac{P_i P_j}{(P_i + S_{ij})(P_i + P_j + S_{ij})}, \tag{8}$$
where P_i is the population of region i and S_{ij} is the overall population inside the circle of radius d_{ij} centered at i.
- Modified Gravity Model [4]: A modified version of the original Gravity Model with two power exponents added. It assumes that the OD flow between two regions is in proportion to the propulsiveness (outflow) and attractiveness (inflow) of the two regions and inversely proportional to the distance between them.
$$f_{ij} = K \frac{{p_i}^{\alpha} {a_j}^{\beta}}{{d_{ij}}^{\gamma}} \tag{9}$$

Traditional Machine Learning Method:

- Random Forest [15]: A tree-based ensemble learning method.

Deep Learning-Based Non-graph Methods:

- Gravity Neural Network [22]: A neural network that takes the same input as our model.
- DeepGravity Model [30]: Our backbone OD prediction model. The brief introduction of DeepGravity can be referred to in Section 3.1.

Statistics	Changsha Partial	Changsha Global	Beijing Partial
City	Changsha	Changsha	Beijing
Districts in training data	Tianxin, Yuhua	Tianxin, Yuhua, Kaifu, Yuelu, Furong	Chaoyang
Districts in testing data	Wangcheng, Changsha County	Wangcheng, Changsha County	Tongzhou
#regions in training data	287	688	510
#regions in testing data	457	457	481
Dimension of attributes	35	35	35
Dimension of treatments	12	12	12

Table 1: The basic statistics of our datasets.

Deep Learning-Based Graph Methods:

- SI-GCN [37]: This model constructs a graph based on the adjacency matrix of regions and then adopts a graph convolutional network to model the spatial interaction. We adapt this model to fit our scenario, with the predictor predicting future OD flow using the concatenation of both the learned representations of O and D and the historical OD flow.
- GMEL [19]: This model constructs a graph of adjacency and adopts two graph neural networks with attention (GAT) to learn the origin embedding and destination embedding, respectively. Besides the OD flow prediction task, the author introduces two downstream tasks to predict the inflow and outflow. The learned embedding of Os and Ds and the historical OD flow are fed into a gradient boosting machine to predict the future OD flow.

4.2 Evaluation Metrics

To measure the prediction performance, we adopt two commonly used evaluation metrics: Common Part of Commuters(CPC) and Mean Absolute Error(MAE). CPC measures the common part of agreements between predicted OD flows and the true OD flows, and it is a widely acknowledged metric for OD prediction. MAE is an intuitive and concise metric for regression problems.

$$CPC = \frac{2\sum_{i,j} min(f_{ij}, \hat{f}_{ij})}{\sum_{i,j} f_{ij} + \sum_{i,j} \hat{f}_{ij}} \tag{10}$$

$$MAE = \frac{1}{|f|} \sum_{i,j} |\hat{f}_{ij} - f_{ij}| \tag{11}$$

where $\hat{f}_{ij}$ denotes the predicted OD flow value for OD pair (i, j) and f_{ij} denotes the corresponding true OD flow value.

4.3 Overall Performance Comparison

To validate our model, we conduct extensive experiments on three datasets and compare the model performance with SOTA baseline methods. The performances of existing methods are listed in Table 2. We can draw four insights from it.

- **Our model's superior performance on all the three datasets:** As we can see from the Table, our model outperforms all SOTA baseline methods across all three datasets. Compared with the best existing method, which is the DeepGravity method, our model achieves performance gain significantly, with a relative gain of 4.25%, 5.60%, and 1.34% on the CPC metric and 10.29%, 12.99%, 5.50% on the MAE metric. Such an improvement greatly proves the validity of our model designs.
- **Our model's robustness across different cities and regions:** Comparing our performance gain for all three datasets, we can conclude that our model achieves robust and consistent performance gain regardless of the city from which the dataset is collected and the scale of the dataset.
- **Insufficient ability of traditional urban planning methods to utilize data.** The traditional urban planning methods, ranging from the gravity model and modified gravity model to the radiation model, all have significant restrictions on the type of data input and only utilize some basic data like population (inflow, outflow) and distance. In this way, these models fail to consider the heterogeneity of regions and therefore only show poor performance.
- **Existing deep learning methods' poor ability to generalize.** Existing deep learning methods fail to generalize to urban planning scenarios where planned areas may have some attribute distribution shift with known areas since they are typically based on correlation to make predictions. Such a spurious correlation is not robust and may not generalize to the testing set, resulting in relatively poor performance. Also, for graph-based methods, since the planned area and known area are two graphs that barely overlap, the graph representation networks are greatly challenged by being transferred to a new graph. This explains the poor performance of graph-based deep learning methods.

4.4 Ablation Study

To further validate every module in our model, we conduct an ablation study on our model, and the results are shown in Figure 6. To show the effectiveness of the decomposed framework and the re-weighting module, respectively, we compare the original OD backbone model DeepGravity(DG) with the corresponding variants, DG+Disentangled Representation module and DG+Re-weighting module. For DG+Re-weighting, the confounding variables we calculate weights on are all the input attributes. Then we compare our final model design, which is the DG+SIRI, with all the models mentioned above. As we can see from the figures, adding either the decomposed framework or the re-weighting module to the original DeepGravity model brings performance gains across all three datasets. Combining the two modules together brings the greatest performance improvement. These results prove the validity of both modules.

Groups	Models	Changsha Partial		Changsha Global		Beijing Partial	
		CPC	MAE	CPC	MAE	CPC	MAE
Traditional Model-based Methods	Radiation Model	0.204	201.60	0.223	137.22	0.280	60.20
	Modified Gravity Model	0.406	56.16	0.388	42.99	0.367	32.74
Traditional Maching Learning Method	Random Forest	0.470	104.40	0.570	64.20	0.730	24.32
Graph-based Deep Learning Methods	SI-GCN	0.270	121.80	0.270	81.31	0.690	37.40
	GMEL	0.537	103.90	0.450	74.22	0.710	45.37
Non-graph Deep Learning Methods	Gravity Neural Network	0.305	114.70	0.457	71.99	0.731	26.52
	DeepGravity(DG)	0.697	39.36	0.697	28.26	0.745	17.28
Our method	**Our Model**	0.728	35.31	0.736	24.59	0.755	16.33
	Performance Gain	4.25%	10.29%	5.60%	12.99%	1.34%	5.50%

Table 2: Experimental results on our datasets.

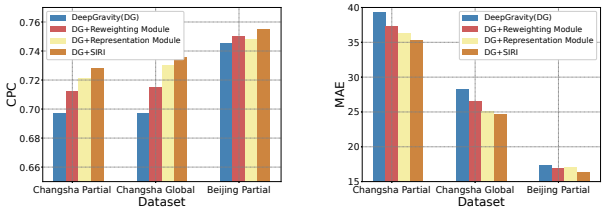

Figure 6: Ablation study results with CPC metric and MAE metric.

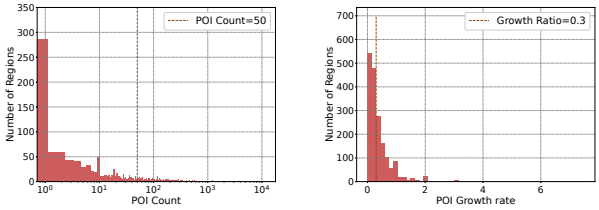

Figure 7: Distribution of POI count and POI growth rate. We define the thirty percentile as the threshold point.

To further analyze the two modules, we compare the performance gain brought by the two modules across all three datasets. As we can see, for the Changsha Partial and Changsha Global datasets, the performance gain brought by the disentangled representation module is slightly greater. For the Beijing dataset, the performance gains brought by the two modules are close. These findings provide evidence for the necessity of solving the OOD dilemma, which our disentangled representation module targets. Meanwhile, the re-weighting module brings a steady improvement to all three datasets.

4.5 Performance in Different Urban Planning Scenarios

In previous sections, we validate the effectiveness of our model in a hybrid urban planning scenario. Despite the overall performance gain, we are still unsure whether our model works for all kinds of specific urban planning scenarios. Therefore, in this section, we do sub-experiments on four typical planning scenarios to test whether the performance gain is consistent and robust. The four typical scenarios are to plan small or large development in less developed or developed regions. We evaluate the relative performance gain of predicted OD flows to regions from different scenarios, which is the ratio of reduced MAE to the mean MAE of the flows. We measure the initial development level of regions with POI count and set the fifty per cent point at 50, as shown in Figure 7, as the threshold of being developed. To quantify the level of planned development, we refer to the POI growth rate and regard 0.3 as the threshold. Our model's performance gains compared with the best baseline method for the four scenarios are shown in Figure 8.

As we can see, our model generates significant performance gains in all four urban planning scenarios, with the lowest gain exceeding 4.5%. These results prove the robustness of our model. Also, for regions that have little planned development, our model

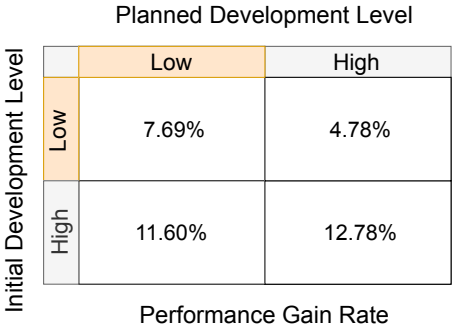

Figure 8: Performance gain rate in MAE metric for four typical urban planning scenarios.

brings great performance gain regardless of whether they are developed. This is within our expectations since such regions are often common in training datasets, and we have more prior knowledge of them. Predicting scenarios that have large planned development is usually more challenging since such cases are relatively rare, and the development of such regions varies mutually. Experimental results show that our model also makes a 4.78% improvement in such scenarios, which proves our method's capacity to precisely model how planned development impacts OD flows.

4.6 Performance Highlights in Predicting Large OD Flows

Previous correlation-based methods often show limited performance in predicting large OD flows since such samples are too rare to learn correlation. However, large OD flows are important and must be paid extra attention to in city governance when involving a large population. For example, a regular large commuting flow between two regions may cause severe traffic congestion and trouble these commuters [36]. To our delight, our model exhibits

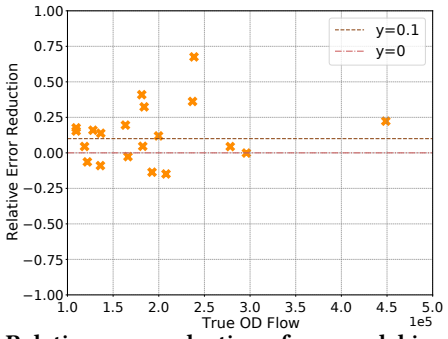

Figure 9: Relative error reduction of our model in predicting large OD flows.

a great performance improvement in predicting large OD flows, as shown in Figure 9. To fairly measure how much performance gain our model brings to an OD pair, we refer to the relative error reduction metric, which is the ratio of prediction error reduction by our method to the true value of our prediction target. As we can see, for OD flows that exceed 10^5, the relative error reduction is significant. 75% of the OD pairs receive a performance gain, and 73% of them receive a relative gain higher than 10%. Although some of the OD pairs receive decreased performance, the decrease is quite slight compared to the overall progress. Such progress in predicting large OD flows illustrates our causal designs' capability to learn how planned development impacts OD flows and then utilize the learned relationship to make inferences in predicting large OD flows. This gives further proof of the value of our model.

In conclusion, we conduct extensive experiments on three real-world datasets and compare our model with four groups of SOTA baselines. Experimental results validate the superiority of our model. We design further experiments to demonstrate the effectiveness of every module and the robustness of our model across all urban planning scenarios. Experiments also show our model's outstanding performance in predicting large OD flows, which are valuable for city governance. These findings corroborate our model's value to the urban planning field.

5 RELATED WORK

OD Prediction: There are typically two kinds of OD prediction problems. The first type is to generate OD flow with contemporary region attributes, and the second is to forecast future OD based on historical OD flow series. Although, in our problem, current OD flow is provided to predict the future OD flow, our task is within the scope of the first type of OD prediction since no historical series are involved. Early researchers solve the first type of OD prediction problem using some model-based methods [4, 38]. Gravity model [18, 39] assumes that the OD flows from one region to another is proportional to the product of both regions' populations and inversely proportional to the distance between the two regions. Radiation model [31] assumes that the probability of a flow departing from region i to region j is intervened by the attractiveness of regions between i and j. With the development of deep learning, data-driven methods are drawing increasing attention [22, 24, 30]. Since OD flows not only depend on origins and destinations but also are affected by adjacent regions, graph representation tools are introduced to grasp spatial influence [10, 19, 29, 35, 37]. These methods typically construct an adjacency graph and adopt GAT [33] or GCN [17] to learn node representations. Then a tree-based regressor or a neural network predicts the OD flow from the learned representations. However, existing OD prediction methods are all correlation-based, making them fail to generalize to out-of-distribution scenarios common in urban planning practice.

Counterfactual Regression: Counterfactual regression problems answer a question: what would the patient have been had he received some treatment? To answer that question requires precise knowledge of the causal relationship between the treatment and the 'patient'. However, since the treatment is assigned to the patient based on one's conditions, there usually exists a correlation between treatment and other features, which brings the confounding bias in estimating the treatment effect. Randomized Controlled Trial [11] is a fair and ideal way to assess the causal effect of the treatment. Since randomized treatment may not be practical in real cases, several methods are proposed to decorrelate treatment and input features statistically. Some proposed re-weighting methods based on propensity [3, 6, 12, 25], and some adopted representation learning methods to decrease the feature representation discrepancy between different treatment groups [16, 27]. Negar et al. [13] moves re-weighting methods forward by finding the essential factors in input features that cause confounding bias and only calculating sample weights on these confounding factors. However, all these methods are initially designed for binary treatment cases and cannot be generalized to bundle treatment. Hao et al. [40] proposed a variational sample weight learning method to deal with bundle treatment cases. Since, in our cases, treatment refers to POI changes of all kinds and is bundle continous treatment, we further generalize [40] to solve our problem.

6 CONCLUSION AND FUTURE WORK

In this paper, we propose a causal decomposition and re-weighting OD prediction method to solve the out-of-distribution problem for OD prediction under urban planning scenarios. We evaluate our method on three real-world datasets from different cities and at different scales. Experimental results validate the effectiveness of model designs. Further analysis of our model shows its robustness and superiority across all typical urban planning scenarios. Overall, our work incorporates causal designs to solve OD prediction tasks in urban planning scenarios, and extensive experiments demonstrate the excellent performance of our designs. In future work, a promising direction is to predict future OD distribution with a larger time interval where longer-term planning takes effect. Also, we may work on more forms of planning other than those involving POIs.

ACKNOWLEDGMENTS

This work was supported in part by the National Key Research and Development Program of China under grant 2020YFB2104005 and the National Natural Science Foundation of China under 61971267, 61972223, and U21B2036.

REFERENCES

[1] David Adams. 2012. *Urban planning and the development process.* Routledge.

[2] Bayes Ahmed. 2012. The traditional four steps transportation modeling using a simplified transport network: A case study of Dhaka City, Bangladesh. *International Journal of Advanced Scientific Engineering and Technological Research* 1, 1 (2012), 19–40.

[3] Peter C Austin. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research* 46, 3 (2011), 399–424.

[4] Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J Ramasco, Filippo Simini, and Marcello Tomasini. 2018. Human mobility: Models and applications. *Physics Reports* 734 (2018), 1–74.

[5] Michael Batty. 2016. Complexity in city systems: Understanding, evolution, and design. In *A planner's encounter with complexity*. Routledge, 99–122.

[6] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. 2013. Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising. *Journal of Machine Learning Research* 14, 11 (2013).

[7] Mingxing Chen, Hua Zhang, Weidong Liu, and Wenzhong Zhang. 2014. The global pattern of urbanization and economic growth: evidence from the last three decades. *PloS one* 9, 8 (2014), e103799.

[8] Peng Cui and Susan Athey. 2022. Stable learning establishes some common ground between causal inference and machine learning. *Nature Machine Intelligence* 4, 2 (2022), 110–115.

[9] Ping Feng, Xiao-Hua Zhou, Qing-Ming Zou, Ming-Yu Fan, and Xiao-Song Li. 2012. Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in medicine* 31, 7 (2012), 681–697.

[10] Yongshun Gong, Zhibin Li, Jian Zhang, Wei Liu, Yu Zheng, and Christina Kirsch. 2018. Network-wide crowd flow prediction of sydney trains via customized online non-negative matrix factorization. In *Proceedings of the 27th ACM international conference on information and knowledge management*. 1243–1252.

[11] Gordon H Guyatt, Jana L Keller, Roman Jaeschke, David Rosenbloom, Jonathan D Adachi, and Michael T Newhouse. 1990. The n-of-1 randomized controlled trial: clinical usefulness: our three-year experience. *Annals of internal medicine* 112, 4 (1990), 293–299.

[12] Negar Hassanpour and Russell Greiner. 2019. CounterFactual Regression with Importance Sampling Weights.. In *IJCAI*. 5880–5887.

[13] Negar Hassanpour and Russell Greiner. 2019. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*.

[14] Lewis D Hopkins. 2001. *Urban development: The logic of making plans*. Vol. 166. Island Press.

[15] Yiqing Huang, Fangzhou Zhu, Mingxuan Yuan, Ke Deng, Yanhua Li, Bing Ni, Wenyuan Dai, Qiang Yang, and Jia Zeng. 2015. Telco churn prediction with big data. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data*. 607–618.

[16] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *International conference on machine learning*. PMLR, 3020–3029.

[17] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[18] Maxime Lenormand, Aleix Bassolas, and José J Ramasco. 2016. Systematic comparison of trip distribution laws and models. *Journal of Transport Geography* 51 (2016), 158–169.

[19] Zhicheng Liu, Fabio Miranda, Weiting Xiong, Junyan Yang, Qiao Wang, and Claudio Silva. 2020. Learning geo-contextual embeddings for commuting flow prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 808–816.

[20] Lester Luborsky, Michalina Fabian, Bernard H Hall, Ernst Ticho, and Gertrude R Ticho. 1958. Treatment variables. *Bulletin of the Menninger Clinic* 22, 4 (1958), 126.

[21] Kevin B Modi, LB Zala, FS Umrigar, and TA Desai. 2011. Transportation planning models: a review. In *National Conference on Recent Trends in Engineering and Technology, Gujarat India*.

[22] Mikhail Mozolin, J-C Thill, and E Lynn Usery. 2000. Trip distribution forecasting with multilayer perceptron neural networks: A critical evaluation. *Transportation Research Part B: Methodological* 34, 1 (2000), 53–73.

[23] Coleman A O'Flaherty. 2018. *Transport planning and traffic engineering*. CRC Press.

[24] Can Rong, Jie Feng, and Yong Li. 2019. Deep learning models for population flow generation from aggregated mobility data. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 1008–1013.

[25] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.

[26] Allen John Scott and Shoukry T Roweis. 1977. Urban planning in theory and practice: a reappraisal. *Environment and planning A* 9, 10 (1977), 1097–1119.

[27] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*. PMLR, 3076–3085.

[28] Zheyan Shen, Peng Cui, Jiashuo Liu, Tong Zhang, Bo Li, and Zhitang Chen. 2020. Stable learning via differentiated variable decorrelation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2185–2193.

[29] Hongzhi Shi, Quanming Yao, Qi Guo, Yaguang Li, Lingyu Zhang, Jieping Ye, Yong Li, and Yan Liu. 2020. Predicting origin-destination flow via multi-perspective graph convolutional network. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 1818–1821.

[30] Filippo Simini, Gianni Barlacchi, Massimilano Luca, and Luca Pappalardo. 2021. A Deep Gravity model for mobility flows generation. *Nature communications* 12, 1 (2021), 1–13.

[31] Filippo Simini, Marta C González, Amos Maritan, and Albert-László Barabási. 2012. A universal model for mobility and migration patterns. *Nature* 484, 7392 (2012), 96–100.

[32] Samuel A Stouffer. 1940. Intervening opportunities: a theory relating mobility and distance. *American sociological review* 5, 6 (1940), 845–867.

[33] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

[34] Ray A Waller and David B Duncan. 1969. A Bayes rule for the symmetric multiple comparisons problem. *J. Amer. Statist. Assoc.* 64, 328 (1969), 1484–1503.

[35] Yuandong Wang, Hongzhi Yin, Hongxu Chen, Tianyu Wo, Jie Xu, and Kai Zheng. 2019. Origin-destination matrix prediction via graph convolution: a new perspective of passenger demand modeling. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1227–1235.

[36] Yanli Wang, Xiaoyu Zhu, Linbo Li, and Bing Wu. 2013. Reasons and countermeasures of traffic congestion under urban land redevelopment. *Procedia-Social and Behavioral Sciences* 96 (2013), 2164–2172.

[37] Xin Yao, Yong Gao, Di Zhu, Ed Manley, Jiaoe Wang, and Yu Liu. 2020. Spatial origin-destination flow imputation using graph convolutional networks. *IEEE Transactions on Intelligent Transportation Systems* 22, 12 (2020), 7474–7484.

[38] Kai Zhao, Sasu Tarkoma, Siyuan Liu, and Huy Vo. 2016. Urban human mobility data mining: An overview. In *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 1911–1920.

[39] George Kingsley Zipf. 1946. The P 1 P 2/D hypothesis: on the intercity movement of persons. *American sociological review* 11, 6 (1946), 677–686.

[40] Hao Zou, Peng Cui, Bo Li, Zheyan Shen, Jianxin Ma, Hongxia Yang, and Yue He. 2020. Counterfactual prediction for bundle treatment. *Advances in Neural Information Processing Systems* 33 (2020), 19705–19715.