**2 Introduction**

**1.1 problem restatement**

As we all know, "drug" has never been a popular word, which not only strongly endangers people's physical and mental health, but also seriously threatens social stability. Once people are exposed to drugs, they will develop strong dependence. After long-term use, they will be more delirious, impulsive, irritable and violent, which will lead to greater medical pressure, rising unemployment, rising crime rate and other negative social effects. Today, the so-called "drug" is not limited to traditional drugs such as heroin, but also includes the abuse of prescription sedatives, which has become a potential crisis in American society. Because opioids still play such a large role in medicine, it is not easy to achieve complete control.

To improve the effectiveness and ability of drug control, it is necessary to have a clear understanding of the mode of drug transmission. Generally speaking, drugs are mainly transmitted in the following three ways: first, through public entertainment places such as bars; second, through introduction among friends and relatives; third, through improper use of opioids in medical facilities. Due to the diversity of transmission modes, the concealment of drug taking process and the addiction of drugs themselves, drugs spread rapidly and covertly just like malaria. For the government, in addition to the huge difficulty in control, drug rehabilitation, treatment and reemployment for drug addicts are also worrying problems. In this regard, we will simulate the transmission mechanism of drugs and explore the source of drugs by tracing the source, so as to provide new ideas for drug control.

众所周知，“毒品”向来不是一个受人欢迎的词汇，它不仅强烈危害人们的身心健康，更严重威胁着社会稳定。人们一旦接触毒品，就会产生强烈的依赖性，长期使用后，更会神志不清、冲动易怒、暴力倾向增加，进而造成医疗压力变大、失业率上升、犯罪率上升等负面社会影响。如今，所谓的“毒品”已不仅限于传统的海洛因等毒品，镇定类处方药物的滥用也包括在内，这已成为美国社会的一大潜在危机。由于阿片类药物在医学上仍然发挥着巨大作用，因此想要实现彻底管控绝非易事。

想要提高毒品管控的效果和能力，首先需要对毒品的传播方式有清晰的了解。一般来说，毒品主要有以下三种传播方式，其一，通过酒吧等公共娱乐场所传播，其二，通过亲朋之间介绍传播，其三，通过医疗场所中阿片类药物的不正当使用传播。由于传播方式的多样性，吸毒过程的隐蔽性以及毒品本身的成瘾性，使毒品如同疟疾一样迅速、隐蔽地传播开来。对于政府来说，除了管控难度巨大，对于吸毒人员的戒毒、治疗和再就业安置同样是令人担忧的问题。对此，我们将模拟毒品的传播机制，通过溯源的方法探求毒品源头，以期为毒品管控提供新的思路。

1.2 Our thinking of the words

We believe that the number of drug users in a region is closely related to some socio-economic factors. From a micro perspective, a person's choice to take drugs mainly depends on the prevalence of drugs around him and his own acceptance of drugs. The higher the prevalence, the greater the exposure to the drug for the individual, and thus the larger the population of drug users. However, an individual's acceptance of drugs is determined by his family background, education level, economic level and other social factors. Therefore, to understand the transmission characteristics of drug abuse behaviors, it is necessary to clarify the relationship between the number of drug users in a certain region and these socio-economic factors, so as to know which region is more prone to drug abuse.

我们认为一个地区的吸毒人数和一些社会经济因素是息息相关的，从微观上考虑，一个人选择吸毒主要取决于周围的毒品流行程度和他本身对于毒品的接受程度。流行程度越高，对于这个个体来说接触到毒品的机会也就越多，进而表现为宏观上的吸毒者数量上升。而个人对于毒品的接受程度则是由他的家庭背景、教育程度、经济水平等多种社会因素共同决定的。所以想要了解吸毒行为的传播特征，就需要理清某地区吸毒人数和这些社会经济因素之间的关系，进而知道什么样的地区更容易毒品泛滥。

**1.3 Our Goals**

**Based on our understanding of the problem and the dataset provioded, we set the following goals:**

**基于对题目的理解和提供的数据，我们设定了以下的目标：**

1. Collate the big data of social economy and drug abuse reports of five states from 2012 to 2016.

对2012到2016年五个州的社会经济和吸毒报告大数据进行整理。

1. Establish a model to describe the relationship between socio-economic factors and drug users in different five states, and find the factors that have the greatest influence on drug users

建立一个模型来描述不同的五个州内社会经济因素和吸毒人数之间的关系，找到对吸毒人数影响最大的要素。

1. Established a new model to simulate the spread of drug abuse behaviors in each county of the five states. And the factors related to the number of drug users found above were introduced as the characteristic parameters of each county. Use the data from 2010 to 2016 to training to find the communication rules that best fit the actual situation.

建立一个新的模型来模拟吸毒行为在五个州每个county的传播，并引入上述找到的和吸毒人数关系较大的因素作为每个县的特征参数。用2010到2016年的数据进行训练，找到最符合实际情况的传播规则。

1. Based on the rules in the existing model and the data in 2010, use the traceability analysis to find the most likely origin of initial drug transmission

基于已有模型中的规则和2010年数据，溯源分析找到最有可能的初期毒品传播原点。

**2 Simplification and Assumptions**

Both the number of drug users and other socio-economic factors need to take into account the size and structure of the population. In reality, the movement and change of population are very complicated. To simplify the reality and take advantage of the value of the model, we make the following assumptions

1. 短时间内政策不会变化，阿片类药物由于医学上的需要仍然会被使用。
2. 执法稳定性。不考虑警察的执法力度差异。当地被抓获的吸毒人数越多，说明当地吸毒人数越多。
3. 无外部性。五个州内的毒品传播只受五个州内以及之间的相互影响，不受其他地区因素的影响。
4. 沟通等价性。任意相邻的county之间沟通紧密度是等价的。

1. The existence of drugs. Opioids will still be used due to medical needs. So laws and policies will not change in the short term,

2. Law enforcement stability. Regardless of the difference in police enforcement. The greater the number of drug users caught locally, the greater the number of drug users locally.

3. No externalities. Drug transmission within five states is affected only by five states and by interactions between them, not by other regional factors.

4. Communication equivalence. Any neighboring county has an equivalent degree of compactness.

3 数据的预处理 Data preprocessing

我们得到了肯塔基州，俄亥俄州，弗吉尼亚州，西弗吉尼亚州和宾夕法尼亚州这五个州共461个county从2010到2016年的一整套丰富详细的社会经济情况的抽样调查报告，包括了当地人的家庭组成、婚姻情况、兵役水平、语言情况、出生地、民族等共149个方面共596条。同时有来自NFLIS的一组这五个州每个县的毒品报告的数据，总计超过20，000条。由于数据量庞大，原始数据中有数据缺失和失真的可能性，会对模型的预测结果产生较大的影响。所以我们需要对于数据进行预处理。

We get a set of rich detailed sampling investigation report on social and economic situation, including the locals of family, marriage situation, service level, language, place of birth, nationality, etc. of Kentucky, Ohio, Virginia, West Virginia and Pennsylvania this five states a total of 461 counties from 2010 to 2016. A total of 149 aspects. There is also data from a group of drug reports from the NFLIS in each of the five states, totaling more than 20,000. Due to the huge amount of data, there is the possibility of data missing and distortion in the original data, which will have a great impact on the prediction results of the model. So we need to preprocess the data.

3.1 缺失的数据

我们将缺失的数据分为以下两种类型：

1. 多年数据空缺。这类数据我们将其直接删除，因为用少量的数据估计出来的数据没有有效性，不能为我们的模型所用。
2. 少量数据缺失。部分数据存在一年或者两年的空缺。这类数据我们通过函数插值方法来完善数据集。

3.1 Default Value Processing

We classified the missing data into the following two types:

A. multi-year data loss. We delete such data directly, because the data estimated with a small amount of data has no validity and cannot be used for our model.

B. a small amount of data is missing. Some data are available for one or two years. Such data we use function interpolation method to improve the data set.

3.2 异常的数据

我们认为异常的数据也会影响模型的预测和模拟。在五个州的社会经济报告数据集中，基于margin of error给出这组数据的置信区间，如果期望值的数据在置信区间中，就接受此数据，否则就拒绝这组数据。

Abnormal Value Processing

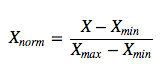
We think that abnormal data will also affect the prediction and simulation of the model. In the socio-economic report data set of five states, the confidence interval of this set of data is given based on margin of error. If the expected value is within the confidence interval, the data is accepted; otherwise, the data is rejected.

3.3 数据的归一化处理

社会经济类的数据集结构复杂，各条数据值之间因为实际意义的不同而差别较大。为了准确地表示社会经济因素对于吸毒行为传播的影响，需要对数据进行归一化处理。我们采用max-min方法（1），将原始数据转化为无量纲的表达式，成为标量。

Normalization of data

The data set structure of socio-economic category is complex, and the data values vary greatly due to the differences in practical significance. In order to accurately represent the influence of socio-economic factors on the spread of drug abuse behaviors, the data need to be normalized. We use the max-min method (1) to transform the original data into a dimensionless expression that becomes a scalar.



3.3 社会类数据的重新分类和整合

原始的数据已经进行了初步的分类，例如经济社会数据分成了17个大类（2010到2012年为16大类，缺少互联网的使用这一类）。并不是所有的类别都是模型所需要的数据，同时部分数据被分类得过于琐碎，不利于模型的简化，所以我们对这十七个类别进行重新的筛选和分类，共形成四个大类：

1. 家庭因素。是当地统计数据中有关不同类型家庭数量的数据，包括household，relation，子女情况，婚姻情况。
2. 教育水平。是描述当地教育资源和居民受教育程度的数据。包括中小学数量，当地居民得学历水平。
3. 特殊人群比例。是描述某协特定的人群在当地的比重，包括军人的数量和残疾人的数量。
4. 文化背景。是描述当地居民整体的文化构成和文化环境的数据，包括出生地，一年前的居住地，语言使用水平，在当地居住的时间以及祖先。

Reclassification and integration of social data

The original data has been preliminarily classified. For example, the economic and social data has been divided into 17 categories (16 categories from 2010 to 2012, lacking the category of Internet use). Not all the categories are the data required by the model. At the same time, some data are classified too trivial, which is not conducive to the simplification of the model. Therefore, we re-screen and classify these 17 categories, forming four categories in total:

1. family factors. Is the data about the number of different types of families in local statistics, including household, relation, children, marriage and etc.

B. education level. Data describing local educational resources and educational level of residents. Including the number of primary and secondary schools, local residents have educational level.

C. proportion of special population. It describes the proportion of a specific population in the local area, including the number of soldiers and the number of disabled people.

D. cultural background. Data describing the overall cultural composition and cultural environment of local residents, including place of birth, place of residence one year ago, level of language use, time spent living there and ancestors.

4 毒品种类

基于NFLIS给出的五个州中每个县的吸毒报告情况，我们对不同种药品的报告数量进行排序，找到找最常见、最容易被滥用的药品。

