Demographical Analysis of US Homelessness and Predictions Based on Long Short-Term Memory Model

Yi Han [1, †], Ding Jin [2, †], Shuyuan Luo [3, *, †]

[1]Department of Undergraduate Studies, Duke Kunshan University, Suzhou, China

[2]School of Information Science, University of Illinois Urbana-Champaign, Champaign, United States

[3]SJTU Paris Elite Institute of Technology, Shanghai Jiao Tong University, Shanghai, China

[*]Corresponding author: luo_sy@sjtu.edu.cn

[†]These authors contributed equally.

Abstract. Homelessness is fast becoming one of the most troublesome issues in United States (U.S.) in terms of the uprising total number of homeless population and potential social instability that it may cause. In previous studies, efforts mainly focused on the characteristics of specific subgroups out of the overall homeless. In this paper, more attention is paid to determine major factors related to homelessness in U.S. using correlation and a model is proposed to predict the tendency of change in the total number for homeless population for each state in U.S. by applying advanced machine learning method of Long Short-Term Memory (LSTM). The study suggests that each state would be primarily troubled by homeless subgroups with different characteristics. Its results also imply that crime rates may relate to with the trend of change in total homeless population and the dominant concerns regarding to the categories of crime differ on a state basis. The study further suggests that after optimization, the LSTM model with interpolation or with multi-dimensions are the best approaches to predict the overall change in the number of homeless populations for each state in the U.S.

Keywords: US Homelessness Prediction, LSTM, Machine Learning.

1. Introduction

Homelessness is increasingly recognized as a serious concern around societies worldwide, particularly in the U.S. Detailing the sights to situations in U.S., the problem demonstrates an uprising tendency through recent years. Beyond the commonly recognized characteristic of homelessness that refers to the roofless situations when people are short of stable and safe housing, its definition expanded to the sense of lacking security or satisfaction of staying at home [1]. Until 2016, there was approximately 550,000 homeless population in U.S. in total [2]. With more and more people experiencing homelessness, it may result in numerous negative social influences, such as huge government costs for hospitalization or an increasing crime rate. For example, according to Youngmin and Andrew, the homeless individuals being victimized provides a potential point where the criminal justice system may intervene with homeless individuals [3]. As the increasing homeless population may also bring an increasing crime rate, it is important to study more on this topic. In this paper, it will focus on the case of U.S. homelessness, analyzing the extent of different categories of crime to affect homelessness and using the approach of LSTM to build a prediction model for overall homeless population from 2016 to 2020.

Reflecting on previous studies, contributions were mainly concentrated on studying specific subgroups from the overall homeless population within rather limited areas. There have been studies which concluded that there are forces to stimulate geographical dispersion of homelessness in the US in terms of metropolitan and urban portions and the history skid districts in selected metro communities [4], and others also observed some annoying higher homelessness rates around urban than rural areas as was estimated using the 2007 homelessness counts and U.S. [5]. Hence, chances are that further studies need to be conducted to decide whether there are other factors influencing the distribution of homelessness across metropolitan areas with respect to the comfortableness to live

outside which is assumed to be determined by climate and overall urban security status by examining crime rates. Moreover, on account of the hope to provide valid insights for policymakers, a prediction model constructed based on the mechanism of machine learning would be favored to give out the tendency of change of the homeless population and specific attention to be paid to subgroups of the homeless. Although there has been research to identify predictors of re-entry and long-term length-of-stay of homeless families within restricted shelters in New York [6] and models for potential risk of becoming chronic homelessness in six months [7] using machine learning, there are still possibilities to apply other machine learning methods to generate a more thorough prediction model regarding to the trend of change in homeless population.

Hence, the major difference between existing research and this paper lies in the advanced machine learning method to construct a prediction model. In this paper, to study the factors that influence homeless population distribution, more diverse factors are considered, such as crime rate and geographic location from 2007 – 2016. As for the prediction model, the method of Long Short-Term Memory in the Recurrent neural network is implemented to predict the trend of the total homeless population. The LSTM algorithm demonstrates advantages in involving time series and bringing flexibility in controlling the outputs with the help of memories of critical ups and downs of certain trends. Therefore, the following part of the paper will discuss the methodology to explain the relatedness of certain aspects and homelessness and detail the mechanism for designing the prediction model.

The remaining part of the paper will be organized as follows: Methodology section summarizes our method, which introduces the technique of analysis. Results and Discission section will be based on a methodology to discuss our results. And it will analyze some possible causes of U.S. homelessness and predict the U.S. total homeless population in a specific period. It also includes the conclusion of our work. And the last section is the abstract of some peer-reviewed articles.

2. Methodology

2.1 Dataset description and preprocessing

The first dataset is “usa-2007-2016-homlessness.csv”, and the source comes from the website [8]. Overall, in this dataset there are 86529 rows and 6 columns, which means it contains six features. Here is the list of six features: Year, State, CoC Number, CoC Name (place where reported), Measures (different types of homelessness). In total there are 42 different measures.

Table 1. Some examples of the various measures

Number	Measures
1	Chronically Homeless Individuals
2	Homeless Individuals
3	Homeless People in Families
4	Sheltered Chronically Homeless Individuals
5	Sheltered Homeless Individuals
6	Sheltered Homeless People in Families

The second dataset is “Crime in Context, 1975-2015.csv”, it comes from the website [9]. Overall, in this dataset, there are 2829 rows × 15 columns, which means it contains 15 features. Here are the 15 features: report year, agency code, population, violent crimes, homicides, rapes, assaults, robberies, months reported, crimes per capita, homicides per capita, rapes per capita, assaults per capita, and robberies per capita. And Table 2 is the description of each feature in this dataset.

Table 2. Some examples of the various measures

Feature name	Description
state	State name
violent crimes	Cases of violent crimes in the given state at the given year
homicides	Cases of homicides in the given state at the given year

rapes	Cases of rapes in the given state at the given year
assaults	Cases of assaults in the given state at the given year
robberies	Cases of robberies in the given state at the given year
population	Population of the given state at the given year
report year	Year name

2.2 The determination of the number of network layers

Missing values for every column need to be checked for both datasets. There are several places of missing data in both datasets. For example, in the first dataset "the Homelessness in USA from 2007 to 2016", there is no figure for the population of "Unsheltered Parenting Youth Age 18-24" in any state in 2007. In this case, mark all the missing values as NA values and then drop all NA values. It also needs to pay attention that since we drop all missing values so it will impact our future analysis. And then move to the next step, which is selecting the feature variables. Since every state is analyzed separately, to make the population of a group of homeless (e.g., sheltered homeless) in one entire state, it is necessary to add up the rows of data that have the same time, same year, and the same state name. (e.g., sheltered homeless AK=AK_500+AK_501). Then all figures for different groups of homeless in each year of one given state are calculated.

2.3 Correlation Analysis

To examine the influence of different categories of crimes rates to overall homelessness, we turn to correlation analysis, aiming to determine the relationship between the crime situation and the overall homeless population. We use the following mathematical formula:

$$r = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}} \quad (1)$$

where $\{x_i, y_i\}, i \in \{1,2,\dots,N\}$ indicates two different subgroups of homeless people, $\bar{x}, \bar{y}$ stand for the mean values of these subgroups, to calculate the correlation coefficients between different types of homelessness and total homelessness for each state, ranking the top types of homelessness and visualizing the results.

2.4 Building prediction models with deep learning algorithms

First, aiming at demonstrating the relationship between the crime rate and the overall homeless population, we use the method of correlation analysis to determine the how different categories of crimes would affect the homelessness. More specifically, it will rank the correlation and visualize the result. Moreover, to predict the change of the total homeless population, using the method of LSTM, in terms of fitting the correlation results to establish a prediction model for different homeless subgroups. By comparing other popular machine learning prediction methods, such as linear regression, and logistic regression, this research ends up manipulating the LSTM approach. This is because the prediction model targets potential change in the number of homeless populations for each state, and the LSTM gains linear units along with a recurrent self-connection through its memory block, which is able to contain multiple memory cells and a pair of gating units for input and output to all cells in the block [10].To prove that LSTM best fits the prediction needs, we first ran the commonly used linear regression to predict population changes. However, this method generated discontinuous curves, which contradict the reality that there was no contingency resulting in shocks in the number of homeless populations during those periods. It thus failed to fit in the requirement of this prediction. After turning to the approach of LSTM, we first performed in the most basic way without interpolation. It provided fundamental shapes of curves demonstrating the tendency. To enhance the accuracy of the predictions, we came up with modifications for the LSTM method from two directions. The first way aims to solve the problem considering the limited size of the original data, the next step was to process the data through interpolation under the framework of LSTM. Nevertheless, since we found that all the predictions were made entirely based on the number of total homeless people, it is also possible to improve is by weighting the changes of the components that

display various significance for different states. Therefore, the other way is to use the multi-dimensional LSTM model. This approach increases both the confidence and accuracy of the results [11]. It obtains the advantage of manipulating the correlation results, which counted the number of the top four correlated subgroups of homeless people and the total homeless population its features.

To build a more suitable making predictions for this research, the experiment starts from proving the failure of simple linear regression model before moving to developing the LSTM method. To construct an appropriate prediction model, we start from training and testing effectiveness by processing current data. In testing the linear regression model, we chose total homeless without manually setting other parameters for each state in order to forecast future number of total homeless population. After normalization the data, we also generate a score to judge the accuracy of the model. Next, as for the LSTM approach, we begin the test by focusing on the number of total homeless as well. Aiming at optimizing mean square error, we set an epoch of 30 and a batch size of 3 to build the LSTM network and use the former 2 data to predict the following data and generate the dataset for testing.

Based on the works above, we would get a basic model to pass the test and optimize it for predictions. Firstly, to optimize the LSTM model with interpolation, the basic structure is the same as that in the LSTM model without interpolation. We expanded the dataset to 500 points for each state with the hope of enriching the results. in this way, we trained the model with the former 400 points and test it with the following 100 points and use the generate the prediction based on the former 30 points. Secondly, to optimize the LSTM on multi-dimension, we normalized the data and reframed the dataset to involve the top four subgroups of homeless population based on the correlation coefficients. To test whether the model works well, we would judge it referring to the root mean square error, with smaller scores standing for better modeling.

3. Results

3.1 Results for Correlation Analysis

We calculated the correlation coefficient between the total homeless population and the correlation coefficient of different homeless subgroups as well. As is shown in Table 3, we extracted results of 8 subgroups. It can be observed that the relationship between different subgroups and total homeless population vary, and it may implicate that different homeless subgroup would exert different impact on the overall situation.

Table 3. The correlation result between total homeless and subgroups in California.

Subgroups of homeless	Correlation between total homeless
Homeless Individuals	0.9735
Chronically Homeless Individuals	0.9699
Unsheltered Chronically Homeless Individuals	0.8975
Unsheltered Homeless	0.8806
Unsheltered Homeless People in Families	0.8734
Homeless People in Families	0.7798
Unsheltered Homeless Individuals	0.7770

Moreover, we also evaluated the relationship between the total homeless population change and different crimes rates by calculated the correlation coefficients for every state to get a more thorough picture. For instance, in Table 4, we show the case in California, some types of crimes, such as the rate of rape, may be negatively related to the homelessness, while other crime rates may be positively to the overall situation.

Table 4. The correlation between total homeless and crime cases in California

Crime name	Correlation between total homeless
violent_crimes	0.8353
homicides	0.9172

rapes	-0.2884
assaults	0.6994
robberies	0.9458

We also demonstrate the results of the top 4 most relevant populations of homeless subgroups and the total homeless population of California here as well as the crime counts and total homeless population as a representative for the result in terms of the situations in every other state in Figure 1 and Figure 2.

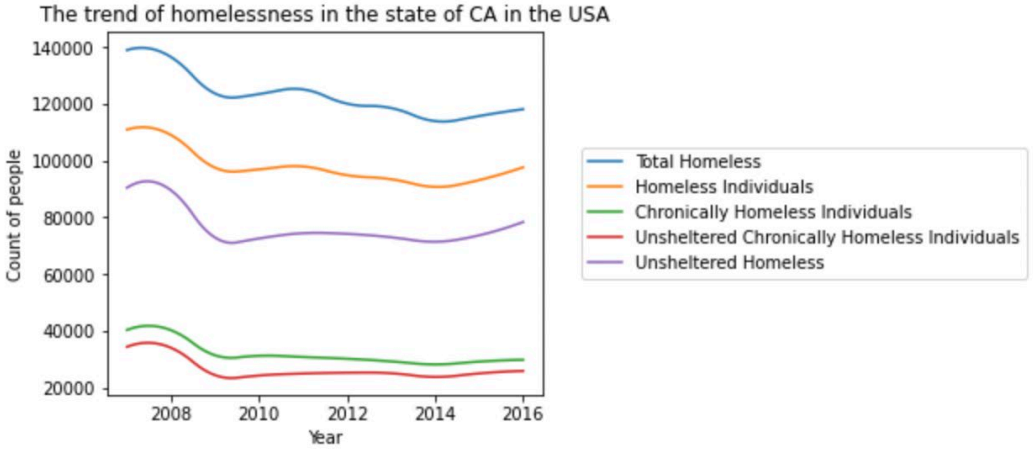

Figure 1. The trend of total homeless and top 4 correlated subgroups in California

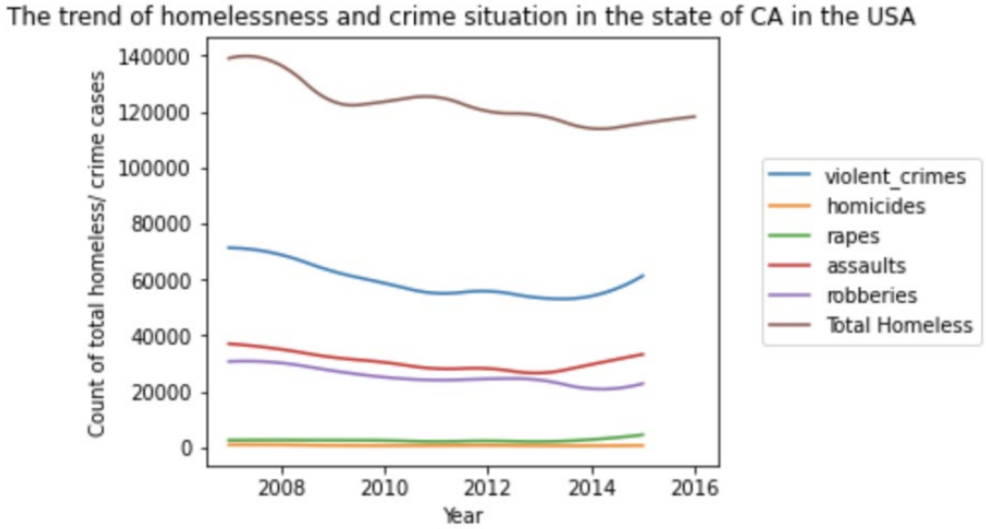

Figure 2. The trend of total homeless and the crime situation in California

3.2 Optimization of LSTM prediction model

To begin with, the linear regression was proved to be inappropriate by reviewing the scores for each state. Although some visualizations for each state seem to be reasonable, after analyzing the scores of fitness, most of them may be considered to be rather too low to prove that the model can successfully stand for the forecast, with an average of 0.4494 and a standard deviation of 0.2906. Therefore, the linear regression model fails in the training process.

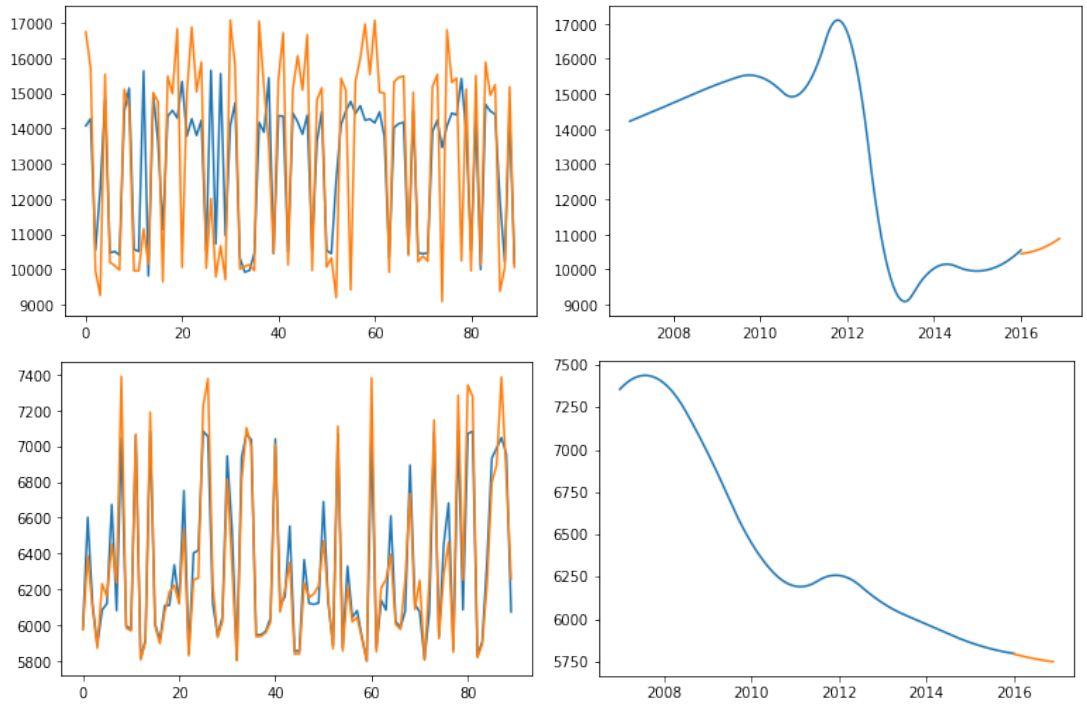

Figure 3. Neural network structure

As is shown in Figure 3, the first 2 graphs on the top reflect situations in AZ while those on the bottom are derived based on the data of IN. The graphs on the left side represents results before normalizing the number of total homeless population, while those on the left are processed after normalization. In all four graphs, the blue lines are drawn with the training dataset, while the orange ones are drawn with the testing dataset. We also calculated the Coefficient of determination (R^2). As is shown in Table 5., among the results of all 54 states, the maximum score is 0.9241 for IN.

Table 5. Coefficient of determination (R2) extracted after training linear regression model

State	Coefficient of determination (R2)
Indiana	0.9241
New York	0.9031
New Hampshire	0.9024
Texas	0.8939
Alaska	0.0209
Wisconsin	0.0168
Rhode Island	0.0088
Delaware	0.0019
(Average score among all states)	0.4494

The LSTM model are proved to be sufficient for making predictions since the test results match the general tendency of change among all the states. In the LSTM model before implementing interpolation, the visualization results appear to be rather general, and it is proved that this method cannot apply for all states. The predictions using this method is shown in Figure 4.

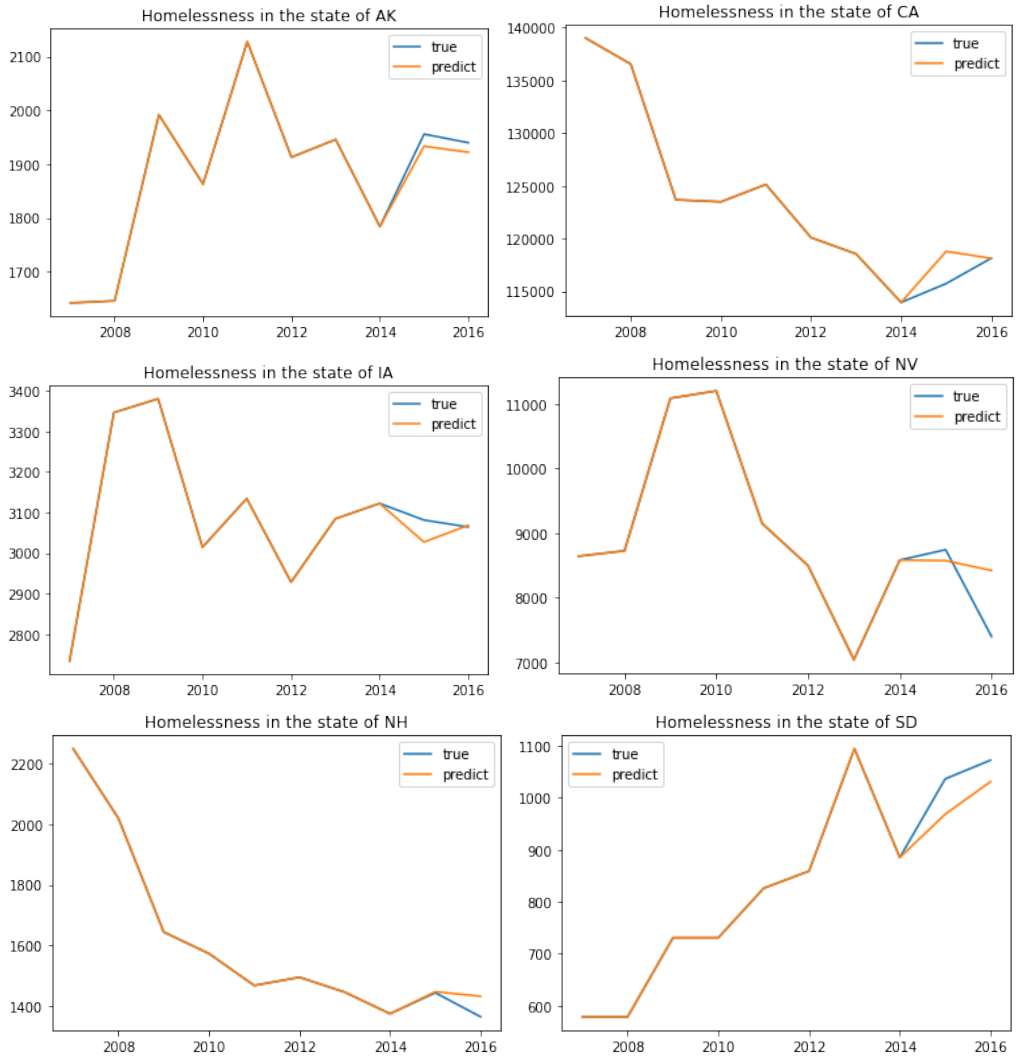

Figure 4. LSTM models without interpolation to predict the change in total homeless population

The graphs demonstrate the LSTM model with interpolation for states in a sequence Alaska (AK), California (CA), Iowa (IA), New Hampshire (NH), Nevada (NV), and South Dakota (ND), from top left to bottom right. The rate monotonic scheduling (RMS) results for LSTM with interpolation approach demonstrate that this model works better compared with that without interpolation. In this case, as is shown in Table 6. the RMS results are rather large is because they were extracted after taking the inverse of normalization of the original data. Among all 54 states, the lowest RMS is derived from IA and the average RMS of all data is 1247.1732.

Table 6. RMS extracted after training LSTM model without interpolation

State	RMS
Iowa	10.9040
Alaska	33.0958
New Hampshire	55.3833
South Dakota	73.8894
California	1785.0686
Michigan	2277.4988
New Jersey	2373.9483
Louisiana	2675.7018
(Average RMS among all states)	1247.1732

After determining the prediction model using LSTM, both optimization approaches, using interpolation or using multi-dimension, are proved to work well in making predictions. Firstly, as for the LSTM model with interpolation, the improvements can be clearly observed as there are more

cases among all states that the prediction parts match well with parts drawn based on the real data. This method gives more reasonable by enriching the original data and generating more smooth curves. The predictions match true data better in the tests, and thus it is more credible in making predictions. In Figure 5, it shows the optimized predictions using LSTM with interpolation.

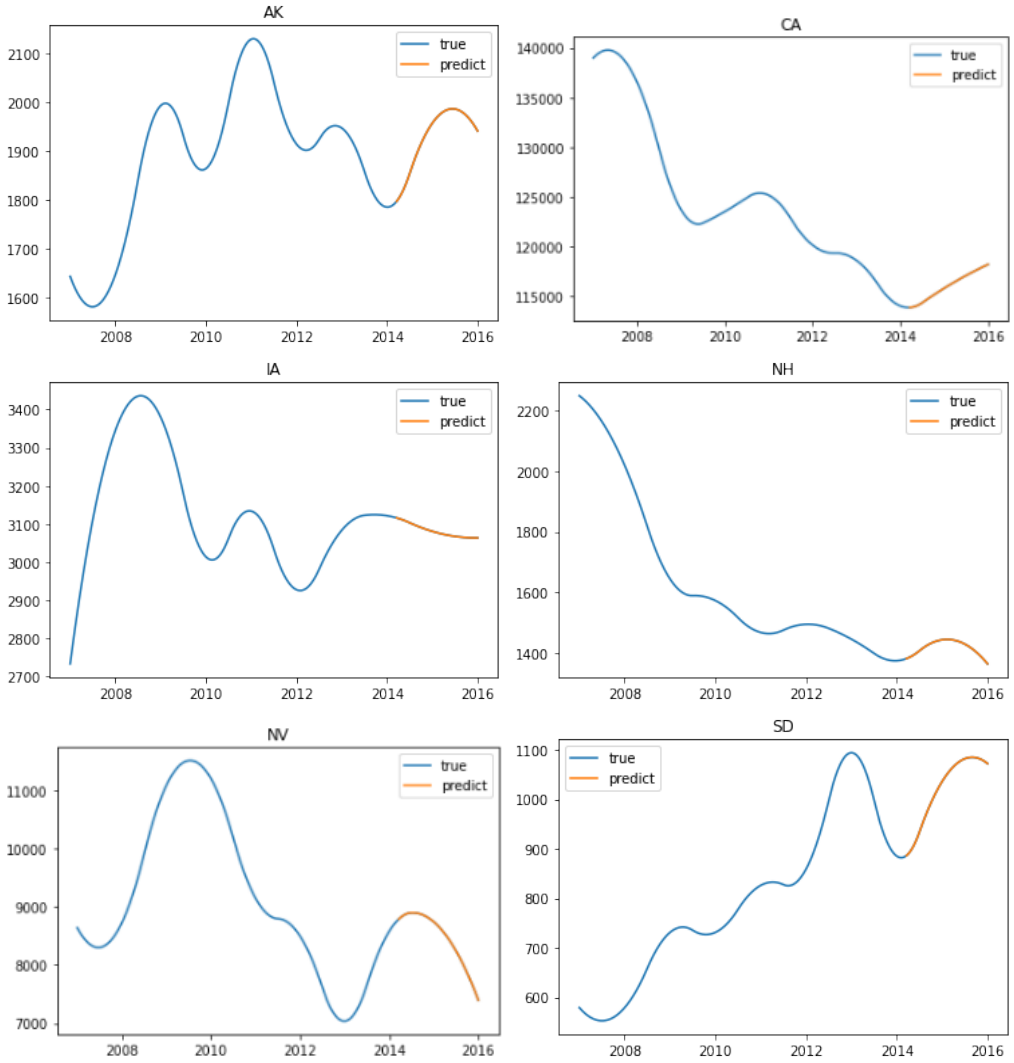

Figure 5. LSTM models with interpolation to predict the change in total homeless population

The graphs demonstrate the LSTM model with interpolation for states in a sequence Alaska (AK), California (CA), Iowa (IA), New Hampshire (NH), Nevada (NV), and South Dakota (ND), from top left to bottom right. The RMS results for LSTM with interpolation approach demonstrate that this model works better compared with that without interpolation as is shown in Table 7. Among all 54 states, the lowest RMS is derived from data of Delaware and the average RMS of all data is 193.8256.

Table 7. RMS extracted after training LSTM model with interpolation

State	RMS
Delaware	0.4883
Wyoming	5.2017
North Dakota	9.6124
New Hampshire	10.5396
California	557.9248
Texas	755.9617
Georgia	1241.1865
Florida	2632.7980
(Average RMS among all states)	193.8256

We also go through the approach of LSTM model using multi-dimensions to predict the trend. It requires to derive distinctive subgroups whose changes are most correlated to the overall homeless population out of the total homeless people from each state. The prediction is as shown in Figure 6. It shows the tendency of change in total homeless population after involving four most correlated subgroups (homeless individuals, chronically homeless individuals, unsheltered chronically homeless individuals, and unsheltered homeless) for CA.

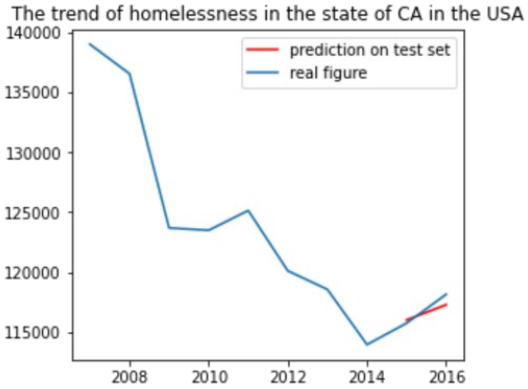

Figure 6. LSTM model with multi-dimension to predict the change in total homeless population

4. Conclusions

In this paper, it analyzed the relationship between different subgroups of overall homeless and between crime and homelessness by calculating correlation coefficients for each state. The paper also developed the LSTM model optimized by interpolation and multi-dimensions to predict the tendency of change in total homeless population in US. In the correlation analysis, it suggested different concerns of for states to pay attention to homeless subgroups with different characteristics. The results also implied that for nearly half of the states in the US, the crime situation may relate to with the trend of change in total homeless population. The relationship between the rate of different types and total homeless population for each state also demonstrate differs. Moreover, to establish the prediction model, after testing the effectiveness of the linear regression model and a LSTM model without interpolation, better approaches are via the LSTM model with interpolation or the LSTM model using most correlated subgroups of overall homeless population as multi-dimensions. The predictions also differ from state to state. However, the limitations are that this research did not arrive at firm prediction on a larger scope in terms of US as an integrated country. In the future, further work may focus on generating predictions for the overall changes in US homelessness, possibly facilitating policy making to enhance social stability and welfare within its border.

References

[1] Van Dam A. The surprising holes our knowledge Americas homeless population [R]. https://www.washingtonpost.com/business/2019/09/18/surprising-holes-our-knowledge-americas-homeless-population/

[2] Kelling K. Older homeless people in London [R]. London: Age Concerns Greater London.

[3] Yoo, Youngmin,Wheeler, Andrew. Using risk terrain modeling to predict homeless related crime in Los Angeles, California [J]. Applied Geography, 2019, 109(4):102039.

[4] Lee B, Price-Spratlen T. The geography of homelessness in American. communities: Concentration or dispersion? [J]. City & Community, 2004, 3(1), 3-27.

[5] Henry Meghan, M. William Sermons. Geography of homelessness [D]. Washington, DC: The Homeless Research Institute at the National Alliance to End Homelessness, 2010.

[6] Hong Boyeong, et al. Applications of machine learning methods to predict readmission and. length-of-stay for homeless families: The case of win shelters in New York city [J]. Journal of Technology in Human Services 2018, 36(1): 89-104.

[7] VanBerlo B, Ross M. A, Rivard, J, Booker R. Interpretable machine learning. approaches to prediction of chronic homelessness. Engineering Applications of Artificial Intelligence [J]. Engineering Applications of Artificial Intelligence, 2021, 102: 104-243.

[8] Vivek Mangipudi. USA 2007-2016 Homelessness. 2016. https://www.kaggle.com/stansilas/usa-2007-2016-homlessness

[9] The Marshall Project. Crime in Context, 1975-2015. 2016. https://www.kaggle.com/datasets/marshallproject/crime-rates

[10] Gers F.A, Schmidhuber J, Cummins, F. Learning to forget: Continual prediction with LSTM [J]. Neural computation, 2000, 12(10), 2451-2471.

[11] Salman A. G., Heryadi Y., Abdurahman E., Suparta W. Single layer & multi-layer. long short-term memory (LSTM) model with intermediate variables for weather forecasting [J]. Procedia Computer Science, 2018, 135: 89-98.