**ANSWER TO REVIEWER COMMENTS**

**TTRA-2018-0169**

The authors would first of all like to thank the reviewers for the positive and constructive feedback on our original manuscript. We believe that your comments have enabled us to more clearly state our scope, results and conclusions and we hope we have reflected this to your satisfaction in our revised manuscript. The responses below address the specific comments. Correspondingly, some changes to the text of the original submission have been made and such changes are marked with red color.

**REVIEWER #1**

1. This paper does not provide very convinced literature review to indicate the background why the CNN model is suitable to forecast the traffic flow, is only that the CNN right now a popular tools? Please provide more convinced relevant literature review to support the very reason why CNN is suitable to be employed to traffic flow forecast.

**Response:**

Thank you for pointing out this issue. We have added more literature reviews on CNN for traffic prediction. Some sentences have been added in the 6th paragraph of section 1 as following:

“Besides, as a representative deep learning method, convolutional neural network (CNN) is widely used in computer vision and image classification (Krizhevsky et al., 2012). In transportation domain, Ma et al. (2017) has demonstrated that CNN is suitable for traffic speed prediction, and after that, Du et al. (2018) uses CNN as a first layer to capture the features of different modality traffic data.”

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

Ma, X., Dai, Z., He, Z., Ma, J., Wang, Y., & Wang, Y. (2017). Learning traffic as images: a deep convolutional neural network for large-scale transportation network speed prediction.*Sensors, 17*(4), 818.

Du, S., Li, T., Gong, X., Yu, Z., & Horng, S. J. (2018). A Hybrid Method for Traffic Flow Forecasting Using Multimodal Deep Learning. *arXiv preprint arXiv:1803.02099.*

1. Section 2.1 should move to the relevant literature review section, due to it divides the traffic flow conditions have temporal and spatial characteristics.

**Response:**

Thank you for pointing out this issue. Section 2.1 had moved to the relevant literature review section in Section 1 paragraph 2 and some of statements have been improved.

1. In addition, in page 6, please provide more statements for the contributions of this paper, at least 150 words for each contribution point.

**Response:**

Thank you for pointing out this issue. In light of the comments of other reviewers, we restated the contributions of this paper as follows:

* A general spatio-temporal feature selection algorithm is proposed.
* A CNN predictive model based on spatio-temporal correlations is proposed.

Deep learning is a general-purpose algorithm based on a large amount of data. When it is applied in a specific field, if it can be combined with the knowledge of the field, it will solve the problem of the domain more accurately and effectively, so that the accuracy of the algorithm can be improved. Thus, on the one hand, the deep learning algorithm needs to be improved to enhance the performance; more importantly, it is necessary to consider the background of the research problem, pre-process the data, and combine this pre-processing with the deep learning algorithm to form a complete algorithm framework to establish a deep learning algorithm suitable for a specific field.

For the first point, as described in Introduction section, the traffic flow always contains rich spatio-temporal characteristics. Therefore, how to effectively use these data is worth studying. In this paper, by verifying the influence of input spatio-temporal data on prediction accuracy, an algorithm for input spatio-temporal data selection is proposed. According to the performance of the prediction algorithm on the verification set, STFSA obtains the input data size that makes the corresponding prediction model obtain high prediction accuracy by searching the input data temporal length and spatial length.It can minimize the error of the model prediction and the effectiveness of the prediction algorithm within an acceptable computation time. The proposed spatio-temporal selection algorithm can be adopted in different traffic scenarios.

For the second point, a CNN predictive model based on spatio-temporal correlations is proposed. Firstly, the spatio-temporal traffic data is transformed into grid data suitable for convolutional neural network processing. In addition, we elaborated on the loss function construction method for noisy traffic flow data. The operation process of the convolutional neural network and the selection of the model hyper-parameters are analyzed, and the detailed parameter configuration in the proposed algorithm is given. Finally, compared to other baseline models, the prediction accuracy of the proposed model turns out to be more efficient by learning spatio-temporal feature.

1. By the way, for the citation problem, line 31 on page 3, Yang and Zhu (1999) is lost in the reference list, please check it carefully. For the reference list problems, (1) please avoid citing working papers, such as Ermagun and Levinson (2016); (2) please also avoid citing meeting conclusion, such as Zhang et al. (2016); (3) please provide complete citation information, such as Kingma and Ba (2014), Wu et al. (2015), and Wu and Tan (2016); (4) please provide relevant DOI for those papers in press status, such as Xu et al. (2018) and Zhang et al. (2018).

**Response:**

Thank you for pointing out this issue. The following references have been updated:

Yang, Z., & Zhu, Z. (1999). A real-time traffic volume prediction model based on the kalman filtering theory. *China Journal of Highway and Transport*, *12*(3), 63-67.

Ermagun, A., & Levinson, D. (2018). Spatiotemporal traffic forecasting: review and proposed directions. *Transport Reviews*, 1-29.

Zhang, W., Zou, Y., Tang, J., Ash, J., & Wang, Y. (2016). Short-term prediction of vehicle waiting queue at ferry terminal based on machine learning method. *Journal of Marine Science and Technology*, *21*(4), 729-741.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Wu, Y., & Tan, H. (2016). Short-term traffic flow forecasting with spatial-temporal correlation in a hybrid deep learning framework. *arXiv preprint arXiv:1612.01022*.

Zhang, W., Qi, Y., Zhou, Z., Biancardo, S. A., & Wang, Y. (2018). Method of speed data fusion based on Bayesian combination algorithm and high-order multi-variable Markov model. *IET Intelligent Transport Systems*, *12*(10), 1312-1321.

We have fixed the aforementioned issues and checked the full text in the revised manuscript.

1. Authors should provide the very details illustrating how the proposed model is working in the experimental results section, i.e., lacking of some essential brief explanation vis-à-vis the text to indicate how the proposed methodology (Figure 1 in page 10) is working in the experimental results section. In addition, for Figure 3 in page 14, please also provide some necessary wordings to guide readers to understand what authors have done and obtained from Figure 3. For Eq. (11) in page 18, please explain why this equation should be involved to be as the loss function; similarly, for Eq. (12) in page 19.  
   Algorithm 1 in pages 26 and 27 is not introduced in text, please provide some necessary illustrations for Algorithm 1.

**Response:**

Thank you for pointing out this issue.

Since we have reorganized the full text, the figure number of the revised manuscript may differ from the original manuscript. Figures 3 in the original manuscript have been changed to Figures 5. We have added the necessary concise text narrative explanations for the figures, tables, etc. in the articles mentioned and marked it in red. For Eq. (11), Eq. (12) and Algorithm 1, we also added more explanatory text and marked it in red in the revised manuscript.

For Figures 1, some explanatory texts have been changed and restated in the 2nd paragraph in Section 2.1.1 as following:

“Figure 1 is a flowchart of the proposed traffic flow prediction framework in this paper. The road traffic data is pre-processed into a matrix form according to temporal and spatial dimensions, as described in Section 2.3.1.The spatial and temporal correlation analysis are conducted in Section 2.1.2. The optimal input data size is determined STFSA, which is given in Sections 2.2. For the prediction method, we use CNN as our prediction model. The inner boxes are the process of determining CNN neural network hyper-parameters that can affect the results of neural network learning and the final regression prediction and they are specified during the training process. The predictive model building process and the specific hyper-parameter configuration are detailed in Section 2.3. The output in the above figure is the predicted output of the model on the test set, which is described in Section 3.”

For Figure 5, some sentences have been added in the 3rd paragraph of section 2.3.1 as following:

“Figure 5 shows a particular input instance construction process. The coil data collected at different locations are integrated into a large two dimensional matrix according to the temporal and the spatial. Each row of the constructed matrix represents the traffic flow of a single detection coil at different times, and each column of the matrix represents the traffic flow of different coils at the same time. The coil detector has a sampling time of 5 minutes.”

For Eq. (11), some sentences have been added in the 2nd paragraph of section 2.3.3 as following:

“where *λ* denotes L2 regularization coefficient, *wj* represents the weight of layers which uses L2 regularization. The L2 regularization method reduces the over-fitting risk of the model by penalizing the large weighting coefficients between the neuron connections, thereby improving the generalization ability of the model.”

For Eq. (12), some sentences have been added in the 2nd paragraph of section 2.3.3 as following:

“The batch calculation error is used to determine the gradient direction of the loss function. In the process of tuning up the weight coefficient of the model using the gradient descent algorithm, the use of mini-batch reduces the influence of a single sample instance with large errors on the entire optimization process by weighting the gradient of a sample set, and speeds up the entire convergence process.”

For algorithm 1, we made some modifications to the algorithm in the original manuscript and updated it in section 2.2. Some sentences have been added as following:

“Table 2 shows the algorithm, and the main idea is to add the most promising data that has not yet been expanded. If we do not found a significant prediction error reduction after *P* expansions of a particular dimension, then we stop searching in that dimension, otherwise we will increase the data length in that dimension. A significant reduction in prediction error means that the prediction error is reduced by more than the insensitive coefficient σ. In the following experiment, *d* was set to 4, *P* was set to six and *δ* was set to 0.1%.”

We also checked other figures, tables and formulas that were not mentioned.

1. For Table 5 in page 31, authors should try to conduct some statistical test to verify the significance of the forecasting performance from the proposed approach. Without the significant test, this paper only has minor contribution. Please refer Diebold and Mariano (1995) and Derrac et al. (2011).  
   F. X. Diebold and R. S. Mariano, “Comparing predictive accuracy,” Journal of Business & Economic Statistics, vol. 13, No. 3, pp. 134-144, 1995.  
   Derrac, J.; García, S.; Molina, D.; Herrera, F. A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. Swarm Evolutionary Computation 2011, 1, 3–18.

**Response:**

Thank you for indicating this issue. We added a two-tailed binomial test (Derrac et al. 2011) in the revised manuscript Section 3.4 and stated in the text as follows:

“Our above analysis dose not answer the question of how much statistical confidence we have in saying our proposed method is “more accurate” than other methods. So we perform a two-tailed binomial test (Derrac et al. 2011) pairwise comparison between the proposed algorithm and other algorithms on the basis of MAPE. An *α*=0.05 level of significance was used for hypothesis testing, and the null hypothesis is that there is no difference in MAPE between two algorithms. The two-tailed binomial test was carried out by the statistical software R. In our experimental framework, a Sign test was performed to compare a total of 32 pairs of comparisons between the STFSA + CNN algorithm and other baseline algorithms. All pairwise comparisons have a p-value of 4.657e-10, much less than 0.05, which means we have at least 95% confidence in rejecting the null hypothesis. The alternative hypothesis suggests that there is a clear distinction between MAPE and the pairwise comparison of other algorithms and proposed algorithm. Furthermore, we performed a Sign test between the ANN and CNN models with or without the STFSA algorithm, and null assumed that the STFSA had no effect on the prediction accuracy of the model. The results show that the p-value is 2.256e-16, so we have at least 95% confidence to reject the null hypothesis and say that STFSA has a gain on prediction accuracy.”

Derrac, J., García, S., Molina, D., & Herrera, F. (2011). A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation*, *1*(1), 3-18.

REVIEWER #2

1. First, the paper is unnecessarily lengthy, and the authors should consolidate the presentations greatly for concisely demonstrating the studies.

**Response:**

Thank you for pointing out this issue. We have simplified the language of the full text, and deleted the redundant part, such as: reduced the detailed description of the comparison model; reduced the unnecessary discussion part; briefed the spatio-temporal correlation analysis part, etc. In order to explain the process of research in more detail, we added the necessary explanation for the diagram and table.

1. Second, the proposed methods and the comparative methods should be given in separate sections for clear presentation.

**Response:**

Thank you for pointing out this issue. A brief introduction to each comparative model and detailed parameter settings have separately given in Section 3.2 and following paragraphs have been added in the revised manuscript:

“Here we describe the experimental details and parameter settings of these models. The python library Keras which is based on Tensorflow is used to build our neural network models. The python library StatsModels is used to build SVR and KNN model. The SARIMA model and statistical tests analysis were deployed on software R. All experiments are performed by a PC Server (the configuration is Intel(R) Xeon(R) CPU E5-2643 3.40GHz, memory 64GB, GPUs is 11G NVIDIA GTX1080TI).

The proposed framework is compared with several baseline traffic flow forecasting methods which are listed follows:

K-Nearest Neighbour, KNN. As a basic nonparametric pattern recognition technique, K-nearest is a widely applied in classification and regression domain (Zhang et al., 2013). An enhanced K-Nearest Neighbor algorithm was adopted as a comparison method. Weighted Euclidean distance, which gives more weight to recent measurements, is used as a similarity measure for KNN (Habtemichael and Cetin, 2016). K is the most import parameter of KNN, which is defined by grid search scheme.

Seasonal Auto-Regressive Integrated Moving Average, SARIMA. A seasonal ARIMA model is formed by including additional seasonal terms in the ARIMA models. The seasonal part of the model consists of terms that are similar to the non-seasonal components of the model, but involve backshifts of the seasonal period. Seasonal ARIMA(p, d, q)(P, D, Q)s model was employed in this paper (Williams and Hoel, 2003) and the Akaike Information Criterion (AIC) is used to determine the appropriate order of ARIMA. All parameters are estimated through maximum likelihood method carried out by the statistical software R.

Artificial Neural Network, ANN. ANNs are a typical neural network model that has undergone significant development in various research fields such as pattern recognition, classification, parameter estimation and prediction. In general, it is composed of three layers of neurons, namely, the input layer, hidden layer, and output layer. The purpose of prediction is to learn and map features of the input data through hidden layer neurons and to generate predictions from the network. In this study, ANN is optimized to contain three layers with 100 hidden neurons. We use ReLU as activation function. Besides, L2 regularization and early stopping is used and the batch size is set to 128.

Support Vector Regression, SVR. Support vector machine is a statistical machine learning algorithm that adopts the structural risk minimization criterion. SVR uses the kernel function to map the low-dimensional nonlinear problem to the high-dimensional space and performs linear problem-solving. The computational process in the transformation space can be implicitly represented by the kernel function. Commonly used kernel functions include the polynomial kernel function, Gauss kernel function, and sigmoid kernel function. The Gauss kernel function is selected for SVR in this study and the parameters of the SVR are determined by a grid search scheme and 10-fold cross validation.

In addition to the SARIMA model, the input data to other models uses spatiotemporal data. The SARIMA model uses single point temporal series data.”

1. Third, computational efficiency is important for online traffic prediction, and this should be investigated in this paper.

**Response:**

Thank you for pointing out this issue. We analyzed the computational efficiency of the proposed model and comparison models, and presented in paragraph 4th of Section 3.4 of the revised manuscript. Some sentences have been added as following:

“Our above analysis dose not answer the question of how much statistical confidence we have in saying our proposed method is “more accurate” than other methods. So we perform a two-tailed binomial test (Derrac et al. 2011) pairwise comparison between the proposed algorithm and other algorithms on the basis of MAPE. An *α*=0.05 level of significance was used for hypothesis testing, and the null hypothesis is that there is no difference in MAPE between two algorithms. The two-tailed binomial test was carried out by the statistical software R. In our experimental framework, a Sign test was performed to compare a total of 32 pairs of comparisons between the STFSA + CNN algorithm and other baseline algorithms. All pairwise comparisons have a p-value of 4.657e-10, much less than 0.05, which means we have at least 95% confidence in rejecting the null hypothesis. The alternative hypothesis suggests that there is a clear distinction between MAPE and the pairwise comparison of other algorithms and proposed algorithm. Furthermore, we performed a Sign test between the ANN and CNN models with or without the STFSA algorithm, and null assumed that the STFSA had no effect on the prediction accuracy of the model. The results show that the p-value is 2.256e-16, so we have at least 95% confidence to reject the null hypothesis and say that STFSA has a gain on prediction accuracy.”

1. Finally, seasonal time series model and k-nearest neighbor model are two conventional short term traffic prediction models, and should be selected in this paper as comparative methods.

**Response:**

Thank you for pointing out this issue.

SARIMA is a linear model that is most widely used in the forecasting field, and KNN is also a representative non-parametric model. Both models are widely used in traffic prediction. We added these two models as a comparison model in the revised submission. The construction process of the model is referred to Williams and Hoel (2003) and Habtemichael and Cetin (2016) and detailed parameter settings have separately given in Section 3.2.

Williams, B. M., & Hoel, L. A. (2003). Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results. *Journal of transportation engineering, 129*(6), 664-672.

Habtemichael, F. G., & Cetin, M. (2016). Short-term traffic flow rate forecasting based on identifying similar traffic patterns. *Transportation Research Part C: Emerging Technologies, 66*, 61-78.

REVIEWER #3

1. The contributions of the work in the context of existing literature is not clear. Presently, the contributions stated are:

*(1) A framework for traffic flow prediction, (2) an optimal input data selection algorithm, and (3) a CNN prediction model based on spatial-temporal correlations.*

(1) & (2) are not contributions to knowledge. (3) can be a contribution provided it is established through literature. So, please rephrase and be specific about the contributions.

**Response:**

Thank you for pointing out this issue. Contributions are updated as below:

* A general spatio-temporal feature selection algorithm is proposed.
* A CNN predictive model based on spatio-temporal correlations is proposed.

We have added more literature support for contribution point 1, and strengthened the theoretical basis and simulation analysis in the choice of spatiotemporal data. We describe the analysis and implementation of the algorithm in detail in Section 2.2 of the revised manuscript. It can be considered as a contribution of this paper.

We added more literature review of first contribution is restated in 2nd paragraph of Section 1 as follows:

“Traffic prediction is a process whereby historical traffic data is used to forecast future traffic situations. Evolutions of traffic conditions are always related to current and past traffic conditions. We hope to identify these relationships and use them for traffic prediction. From the perspective of the time domain, traffic flow data can be viewed as a time series. The different components of a time series dataset, such as its trend and seasonal, cyclical, weekly, monthly and annual variations, can be used as data features. Predictions of future traffic flow data are extend of current data (Tang et al., 2017). In the spatial domain, the traffic flow conditions of a particular road section are closely related to those upstream and downstream of that road section (Park et al., 2011; Wu and Tan, 2016). Since Okutani et al. (1984) first applied the spatial characteristics of road traffic to traffic flow prediction, more and more traffic prediction studies have considered spatial correlations between traffic and road segments (Ermagun and Levinson, 2018; Stathopoulos and Karlaftis, 2003). Spatial correlations, in this case, refer to relationships between upstream and downstream road sections of traffic, and the consideration of such correlations provides multiple references for traffic prediction. Park et al. (1998) found the traffic flow of upstream road sections to be highly correlated with that of the focal section, and reported that spatial information is as informative as temporal information. The current traffic volume of a road section is not only related to the upstream volume, but is also affected by the downstream volume. This situation is particularly evident in the case of traffic congestion (Abdulhai et al., 2002). Therefore, changes in traffic flow conditions have certain temporal and spatial characteristics. There are few articles on the study of spatio-temporal data selection. For example, Yu et al. (2016) have made random combinations of different input spatial data to improve prediction accuracy, Ma et al. (2017) and Polson and Sokolov (2017) made empirically judge the temporal length. How to effectively select input data deserves further study.”

We have added more elaboration of the contributions mentioned in the text to the revised manuscript in 7th, 9th and 10th paragraph in Section 1.

1. The matrix presentation methods described for choosing the temporal and spatial input points seems heuristic /arbitrary. The ACF plot presented is for non-stationary data and 0.68 cannot be arbitrarily chosen as a cut-off! The spatial correlation method described is quite vague. How are the datasets are actually matched? Pearson’s correlation? What about lagged spatial correlation??

**Response:**

Thank you for pointing out this issue. We propose a spatio-temporal feature selection algorithm with respect to a particular prediction algorithm, taking into account its heuristics, biases, and trade-offs. Because, in a practical rode traffic prediction task, there are two problems we must face with: (1) Our data sets is only a sample of real data and the actual data full distribution is not available to the predictive model; (2) There are too many combinations of input features, and most practical algorithms attempt to find a hypothesis distribution by approximating NP-hard optimization problems. Due to above reasons, we propose a spatio-temporal feature selection algorithm with respect to a particular prediction algorithm, taking into account its heuristics, biases, and trade-offs. We added the following explanatory text to the updated version in the 5th and 6th paragraphs:

“In a practical rode traffic prediction task, there are two problems we must face with: (1) Our data sets is only a sample of real data and the actual data full distribution is not available to the predictive model; (2) There are too many combinations of input features, and most practical algorithms attempt to find a hypothesis distribution by approximating NP-hard optimization problems (Kohavi and John, 1997). For the first problem, it is related to the bias-variance trade-off problem: one must trade off estimation for more parameters (bias reduction) with accurately estimating these parameters (variance reduction) (Kohavi and Wolpert, 1996). For the second problem, finding the "optimal" data distribution hypothesis is intractable because fitting the optimal model is NP-hard.

From above analysis, we propose a spatio-temporal feature selection algorithm with respect to a particular prediction algorithm, taking into account its heuristics, biases, and trade-offs. Then, this tuning process is reduced to find an input with high accuracy. ”

This is true, 0.68 cannot be selected as a truncation, and we have re-described the correlation analysis in the revised manuscript. We use the Pearson correlation coefficient to indicate the degree of correlation between different road segments, the cross-correlation analysis in the original manuscript does not reflect the correlation between the different observation points well, so we replaced it with the heap map of Figure 2(b). In our analysis, we do not consider the lagged spatial correlation, but assume that the CNN can handle the lagged spatial correlation, because a lagged spatial correlation between two adjacent observation points is represented in the spatiotemporal traffic flow matrix as two diagonally adjacent elements. The CNN is capable of handling the interrelationship between different relative positions of the input two-dimensional data.

1. In figure one the prediction methods contain a few boxes which do not show any sequence details of the process. It is necessary to provide more information in significantly more details on how the boxes are linked and what happens inside a box. For eg. ‘Determine network structure’ Does not give information about which network, how to structure is determined, what elements of the structures are determined at every step of prediction or during the training stage. It is also unclear whether all these steps are necessary to be followed during the production process or just during the training of the network.

**Response:**

Thank you for pointing out this issue. The inner boxes are the process of determining CNN neural network hyper-parameters that can affect the results of neural network learning and the final regression prediction and they are specified during the training process. The predictive model building process and the specific hyper-parameter configuration are detailed in Section 2.3. We have revised the expression in the original text below Figure 1 in 2nd paragraph of Section 2.1.1, and reiterated it as follows:

“Figure 1 is a flowchart of the proposed traffic flow prediction framework in this paper. The road traffic data is pre-processed into a matrix form according to temporal and spatial dimensions, as described in Section 2.3.1.The spatial and temporal correlation analysis are conducted in Section 2.1.2. The optimal input data size is determined STFSA, which is given in Sections 2.2. For the prediction method, we use CNN as our prediction model. The inner boxes are the process of determining CNN neural network hyper-parameters that can affect the results of neural network learning and the final regression prediction and they are specified during the training process. The predictive model building process and the specific hyper-parameter configuration are detailed in Section 2.3. The output in the above figure is the predicted output of the model on the test set, which is described in Section 3.”

For clarity, we have added more explanatory words in sections 2.3.3 and 2.3.4.

1. How is the initial correlation analysis for matrix presentation is linked with STFSA? Does it provide/generate Rinitial? Why two steps are necessary? Why not start with STFSA directly rather than using arbitrary correlation analysis?

**Response:**

Thank you for pointing out this issue. We reanalyzed and designed the STFSA algorithm and detailed in Section 2.3. The initial input data selection of 0.68 in the original manuscript seems unreasonable, and we have removed this part in the revised manuscript. The STFSA in the revised manuscript can work on its own and no longer needs Rinitial. In Section 2.2, we provide a more detailed example analysis and algorithm design reasons.

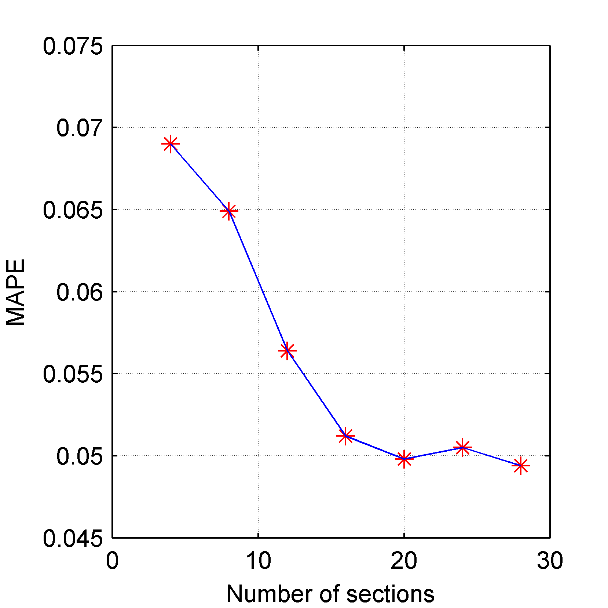
1. Fig 6 presents some confusing results. It shows that with longer length of data the prediction accuracy decreases. This is counterintuitive. The MAPE should show that prediction accuracy increases when input data length increases between 5-20mins and then it levels out for further data points. Again, if the input length in increased significantly to up to 24 hrs or longer, we should see another improvement in prediction accuracy. Longer length of data should not affect the prediction accuracy. For spatial sections, it was initially described only 8 neighboring detectors were used…however the results are presented for nearly up to 30 detectors. Please explain.

**Response:**

Thank you for pointing out this issue. First of all, we should explain that the original manuscript does not clearly explain this part, which may lead to readers' misunderstanding. The conclusion of original manuscript in Figure 6 corresponds to the 5-minute traffic prediction problem. For different prediction problems, if the predicted length changes, data selection and model training should be repeated.

For the accuracy decreases, as can be seen from Figure 3(b) of Chapter 2.2, taking NO 95 observation point as an example, for the 5-minute traffic flow forecast, the extended time lag does not bring about an effective improvement in improving the prediction effect for different spatial length. Further, if the time interval is steadily increased, the prediction accuracy can be a bit reduced. Similar research phenomena can be found in Figure 8 of Du et al (2018). For a short-term traffic flow prediction problem, researchers often empirically choose a relatively short observation time such as Ma et al. (2017) and Polson and Sokolov (2017) for less than 40 minutes.

For other data sets, prolonging the length of the observations to more than 24 hours may result in an increase in accuracy, which may be due to the seasonality of the traffic data. But as Polson and Sokolov (2017) declared in discussion, future traffic conditions are more similar to current ones as compared to those from previous days. Thus, allowed us to develop a powerful model by using recent observations as model features. Besides, our proposed STFSA can extend the prediction time lag up to 24h or even longer if the added data can bring error reduction.



The picture above is Figure 6 in the original text. This is a curve in which the prediction error increases with the number of spatial nodes. We select 8 detection coils in advance in order to compare the prediction error of 8 detection coils with the prediction error after feature selection. However, the selection process of 0.68 is not reasonable, and this part has been removed in the revised manuscript.

Du, S., Li, T., Gong, X., Yu, Z., & Horng, S. J. (2018). A Hybrid Method for Traffic Flow Forecasting Using Multimodal Deep Learning. *arXiv preprint arXiv:1803.02099*.

Ma, X., Dai, Z., He, Z., Ma, J., Wang, Y., & Wang, Y. (2017). Learning traffic as images: a deep convolutional neural network for large-scale transportation network speed prediction. *Sensors*, *17*(4), 818.

Polson, N. G., & Sokolov, V. O. (2017). Deep learning for short-term traffic flow prediction. *Transportation Research Part C: Emerging Technologies*, *79*, 1-17.

1. Apart from the aforementioned issues, the rest of the results seem appropriate. But the STFSA which is claimed to be the main contribution of the work, does not seem to make a major difference to the MAPE error of the ANN or SVM accuracies published elsewhere. I understand that the model accuracy is dependent on the data characteristics to some extent, however STFSA framework is not making any significant difference.

**Response:**

Thank you for pointing out this issue. Firstly, the same data set should be used in the performance comparison of different predictive algorithms. Secondly, it can be seen from Figure 3 in Chapter 2.2 that the temporal length and spatial length does have a large impact on the final prediction error value. We analyzed the short-term traffic flow prediction for 5-20 minutes, and finally we came to an conclusion, for different forecasting tasks, the impact of changes in time and spatial dimensions on the final forecast is unbalanced. For example, we found 5 minutes of traffic in different time and space data organization methods at different intervals of 20 intersections. The main source of influence of the flow prediction process error is the amount of spatial data. However, this situation is not applicable to the 10-15 minute traffic prediction problem, so we finally found that the necessary data selection process for different prediction tasks. It is indeed effective in reducing the error of prediction. Finally, in Section 3 of our revised manuscript, the prediction performance of different algorithms at different observation points for different time periods is analyzed. The statistical test in 4th paragraph of Section 3.4 shows that STFSA can effectively improve the prediction accuracy.

1. The input data size for ANN, SVR and CNN as it is needs to be presented and compared with ANN+STFSA, SVR+STFSA and CNN+STFSA. Computational time and efficiency needs to be discussed.

**Response:**

Thank you for pointing out this issue. We conducted experiments on the traffic flow prediction of multiple road sections to verify the applicability of the proposed method to different road observation points. Since different observation points STFSA may produce different outputs, in Section 3.3, we took the analysis coil of No. 95 as an example to give an analysis process for a specific road section. Parameter configuration, final use data for models and comparison models are detailed and added in the revised manuscript as follows:

“To perform more general test, we enact 5 minute traffic flow prediction task for 8 different observation points of road traffic. In addition to the STFSA model for dynamically determining input data and SARIMA using only single-point temporal series data, other predictive models use a fixed consistent input data with a temporal length of 28 and a spatial length of 12. It means that there are a total of 12 upstream and downstream observation points in the past 140 minutes of traffic flow that are used to predict a single point flow for the next 5 minutes. For example, after STFSA on NO.95 loop detector, the ANN+STFSA model selects an input with a temporal length of 4 and a spatial length of 16, while the CNN+STFSA model selects an input with a temporal length of 4 and a spatial length of 16.The numerical results are reported in Table 6 and Table 7.”

We analyzed the training time and efficiency of the proposed model and comparison models, and presented in Figure 9 in Section 3.4 of the revised manuscript. Moreover, a paragraph has been added in the revised manuscript as follows:

“Figure 9 (b) shows the training time of different models for different forecasting time. Since there is no parameter estimation process, which makes the KNN model the shortest training time. The neural network scale in this paper is small and assisted by the GPU's matrix computing power, making ANN and CNN more efficient than SARIMA and SVR. Besides, the input of multidimensional data increases the computational burden of the SVR and reduces its efficiency. For SARIMA model, it takes about 4 minutes to fit the parameters of the model. For models using STFSA, it usually takes more than ten times of the training time of the non-STFSA model to get an optimized model, which is a trade-offs between prediction accuracy and training efficiency. Compared to other models, the training time of STFSA+CNN is also within an acceptable range. In addition to the SARIMA algorithm, the online prediction time of other models is within 1 second, which is suitable for real-time prediction. SARIMA often needs a rolling fit in order to obtain accurate prediction results, and the fitting time is too long to be suitable for online prediction.”

1. The paper should be organized better. STFSA should be presented before other methods. Also, no need to give so much details of CNN, ANN or SVR. Provide appropriate reference.

**Response:**

Thank you for pointing out this issue. The STFSA algorithm is given separately in section 2.2. We have reduced the detailed description of the comparison algorithm. In addition, the SARIMA and KNN models are added as two baseline models. A brief introduction and detailed parameter settings for each comparison algorithm are given in Section 3.2, and an appropriate reference is provided for each comparison algorithm. Besides, we believe that it is also important to deploy CNN in the transportation field, so the description of CNN is retained.

We optimized the structure of the full text in the revised manuscript, and streamlined the linguistic expression of the full text, removed some of the redundancy, and added more explanations to the figures and tables to guide the reader through the work done by the author.

1. The phrase ‘time correlation’ should be changed to ‘temporal correlation’.

**Response:**

Thank you for pointing out this issue. We checked the full text and corrected the statement.

1. Figure 1: There are some typos in this figure. The word person should be changed to Pearson.

**Response:**

Thank you for pointing out this issue. We checked the full text and corrected the statement.