A Novel Approach to Air Passenger Index Prediction: Based on Mutual Information Principle and Support Vector Regression Blended Model

Honglin Xiong¹,², Chongjun Fan², Hongmin Chen¹, Yun Yang¹, Collins Opoku ANTWI², and Xiaomao Fan³

Abstract

Air passenger traffic prediction is crucial for the effective operation of civil aviation airports. Despite some progress in this field, the prediction accuracy and methods need further improvement. This paper proposes an integrated approach to the prediction of air passenger index as follows. Firstly, the air passenger index is defined and classified by the K-means clustering method. And then, based on mutual information (MI) principle, the information entropy is used to analyze and select the key influencing factors of air passenger travel. By incorporating the MI principle into the support vector regression (SVR) framework, this paper presents an innovative MI-SVR machine learning model used to predict the air passenger index. Finally, the proposed model is validated by air passenger throughput data of the Shanghai Pudong International Airport (PVG), China. The experimental results prove MI-SVR model feasibility and effectiveness by comparing them with conventional methods, such as ARIMA, LSTM, and other machine learning models. Besides, it is shown that the prediction effect of each model could be improved by introducing influencing factors based on MI. The main findings are considered instrumental to the airport operation and air traffic optimization.

Keywords: airport operation and management, air passenger index (API) prediction, machine learning, mutual information, SVR, K-means

Introduction

The global “digital divide” status quo is quickly changing with the progress in artificial intelligence (AI) technologies and their application area expansion. Nowadays, AI has been under researched, the heart of AI technologies is machine learning (ML), which has branched into shallow and deep learning (Gunning & Aha, 2019). Modeling these areas is often an approximation of the objective phenomena involved by machine learning methods. Examples of shallow ML models are the support vector machine (SVM) invented by Cortes and Vapnik (1995), and neural networks (NN). Numerous shallow ML models have been deployed in many research fields, including speech and natural language processing, computer vision, and public opinion mining (Barth et al., 2019; Chen et al., 2020; Guo & Zhang, 2020; Niu et al., 2020; Ullal et al., 2021; Zou et al., 2019). The linear and nonlinear shallow models have gained impressive results in regression and prediction studies, including the prediction of airport passenger throughput with high accuracy (Li, Han, et al., 2018; Li & Jiang, 2020; Sun, Lu et al., 2019). Such linear models as autoregressive integrated moving average (ARIMA; Li, Han, et al., 2018) and gray model (GM; Li, 2014) have achieved significant predictive results. However, nonlinear models, including the BP-neural network (Dantas et al., 2017), recurrent neural network (RNN; Başaran and Ejimogu, 2021), long- and short-term memory network (LSTM; Greff et al., 2017; Kong et al., 2017; Li & Cao, 2018), and support vector regression (SVR; Liang et al., 2015; Li, Ni, et al., 2018; Sun et al., 2011, 2014; Tuo, 2012) were found to be more robust than linear ones due to their strong fault-tolerance levels.

Current airport passenger throughput prediction literature portrays the predominance of research output from mature
western aviation markets. These markets possess high statistical regularity (Sun, Lu, et al., 2019; Sun, Wei, et al., 2019), and therefore, predictive models used may not generally be appropriate for emerging markets. To overcome the statistical irregularity in developing markets, the mutual information (MI) principle of the probability and information theories, which reflects not only linear but also a nonlinear correlation between variables is used in this study for the airport passenger throughput prediction for the PVG, which is a typical international hub airport in mainland China. The superior predictive capacity of blended MI and ML models is discussed in the literature (Chernyshov, 2013; Sharmin et al., 2019). Using the achievements of prior studies, this paper not only focuses on passenger throughput of civil aviation airports but also processes ancillary data related to passenger throughput, to obtain a comprehensive visual representation of air passenger travel. Based on relevant literature, the current study combines the information theory knowledge with the SVR method and calculates MI to enable the selection of key influencing travel characteristics to build the blended MI-SVR model. The main goal of this blended model development is to achieve a better forecast that accurately guides passengers’ choice of travel time, given the selected influencing features of air travel.

Machine learning abilities can take digital marketing to new heights with their AI making all the difference. Of course, ML methods are more and more widely applied to air passenger transportation. In terms of route flow, the gray model is used for route passenger flow prediction (Xia et al., 2016), while classical neural network models and other blended models are exploited for the hybrid prediction of airport passenger and freight route flows (Sulistyowati et al., 2018). The regional aviation passenger flow aspect is covered by the neural networks and SVR models (Sun, Lu et al., 2019), which effectively predict the regional aviation market trends of the short- and long-term passenger flow. The effective strategy, which allows a regional aviation company to increase the hub passenger flow influence in the existing route network, was proposed (Tuo, 2012). Air passenger flow characteristics are quite accurately predicted based on the time sequence columns (Kong et al., 2017; Li & Cao, 2018; Sun et al., 2014; Ullal et al., 2021) and nonlinear vector autoregressive neural network (Li & Jiang, 2020). Jun et al. (2010) stated that comprehensive gray correlation method was adopted to analyze the factors influencing air passenger volume, identify the main ones, and elaborate the respective multiple regression model. As follows from the above brief survey, the attempts to improve the model prediction accuracy in this field involve either a refinement of a particular single model or blending of several models, to achieve better results.

The MI approach can reflect both linear and nonlinear correlations between variables. The efficiency of feature selection based on MI (Kinney & Atwal, 2014; Sharmin et al., 2019) and ML field (Cai et al., 2021; Chernyshov, 2013; Sun, Wei et al., 2019) is well-known. In contrast to the previous studies, based on the analysis above and to resolve the airport passenger forecasting problem, this article contributes to the existing body of literature in two respects: first, a new air passenger index and evaluation method is proposed. Second, a novel air passenger index prediction approach, combining mutual information and original SVR method, will be applied to predict airport air passenger index.

This paper analyzes the air passenger throughput of a civil aviation airport and processes the air passenger throughput as the basic data. Thus, it can directly reflect the travel situation on air passenger flow and provide auxiliary data for the monthly decision-making of civil aviation airport’s operation service. Besides, available methods mostly take into account the accumulated historical data and neglect other related factors that affect the travel of air passengers. Based on the research of other scholars, this paper introduces the knowledge of information entropy and SVR method, selects key influencing factors through the calculation of MI to build an ML prediction model, using this way to find a better prediction method for the travel schedule selection by which passengers choose and provide the relevant reference basis.

The rest of this paper is structured as follows. Section 2 clarifies the research topics and methods to address them; section 3 introduces the MI and SVR original theory, and then presents the blended MI-SVR model. Section 4 analyzes the passenger throughput and weather data of the PVG, to verify the MI-SVR model’s feasibility and effectiveness, in addition, the methodology is used to evaluate in the PVG air passenger index. Section 5 concludes the paper and discusses future research avenues.

### Research Topics

The rising demand for civil aviation airports has created enormous operation and management databases that exhibit significant differences between particular airports. Therefore, it is crucial to identify the contextual factors of particular airport’s service capabilities that would facilitate the provision of required support to airports’ operations and management and furnish solutions for the public air traffic optimization. This implies the necessity to convert the raw data on air passenger throughput to the air passenger index (API), which takes into account the airport’s capacity and variability in passenger traffic volumes. The following definitions provide the mathematical representation of API and the level of air passenger index (LAPI).

**Air passenger index (API):** Setting the value of \(X_i\) as the airport passenger throughput per unit period, the API for this period is defined as \(X^*_i\):

\[
X^*_i = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}}
\]

where \(X_{\min}\) and \(X_{\max}\) are the minimum and maximum numbers of air passengers per unit time, respectively, while \(X^*_i\) varies between 0 and 1.
Level of air passenger index (LAPI): Set \( \{P_1, P_2, \ldots, P_t\} \) is the API sequence set of multiple time units. After clustering, the generated cluster \( \{N_i\} \) is a collection of data objects. The value is the data element of each cluster after clustering. If the API’s unit period is 1 month, and the airport’s monthly API is \( X'_t \), the level of air passenger index (LAPI) is derived as:

\[
N_{X'_t} = \begin{cases} 
1, p_t \in (0, i) \\
2, p_t \in (i, j) \\
3, p_t \in (j, k) \\
\ldots \\
N, p_t \in (\theta, 1) 
\end{cases} \quad (2)
\]

Where \( p_1, p_2, p_3 \) are time units, while \( i, j, k, \theta \) are cluster boundary values.

The API and LAPI indexes of civil aviation airports are instrumental in optimizing their operation and management solutions and improving their services. They are incorporated into the blended MI-SVR model elaborated in this study.

**Related Theory and Model Elaboration**

**Mutual Information Theory**

The information entropy is a key indicator in the information theory, which was introduced by Shannon in 1948 (Fan et al., 2013) based on the following concept: the more ordered is a system, the smaller is its information entropy and vice versa. Therefore, information entropy can be used to measure the degree of system uncertainty (or degree of ordering) (Fan et al., 2013) based on the following concept: the more ordered is a system, the smaller is its information entropy and vice versa.

Information entropy can be derived via the following formula:

\[
H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i) \quad (3)
\]

where \( P(x_i) \) is the probability of sample \( x_i \), and \( n \) is the number of samples. It can be seen that the smaller is the occurrence probability of an event, the higher are the information uncertainty and entropy values.

Let the joint probability distribution of the random vector \( (X, Y) \) be \( P_{xy} \), then the two-dimensional joint entropy of vector \( (X, Y) \) is:

\[
H(X, Y) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} \log p_{ij} \quad (4)
\]

Assuming that the joint probability distributions of \( X \) and \( Y \) are \( P_{x} \) and \( P_{y} \), respectively, the conditional entropy can be defined as:

\[
H(Y \mid X) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} \log \frac{p_{ij}}{p_{y}}
\]

\[
H(X \mid Y) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} \log \frac{p_{ij}}{p_{x}}
\]

\[
H(X, Y) = H(X) - H(Y \mid X) = H(Y) - H(X \mid Y)
\]

Thus, MI can be expressed as an entropy value for which the variable \( X \) (or \( Y \)) is reduced due to the occurrence of the variable \( Y \) (or \( X \)).

\[
I(X; Y) = H(X) - H(X \mid Y) = H(Y) + H(X) - H(X, Y)
\]

By combining formulas (3)–(6), the complete expression of MI can be reduced to the following form:

\[
I(X; Y) = \sum_{i,j} p_{ij} \log_2 \frac{p_{ij}}{p_{i}p_{j}}
\]

**Support Vector Regression (SVR)**

The support vector regression (SVR) is a widely used ML prediction method, which adopts the principle of minimizing structural risk rather than minimizing empirical risk. This allows one to effectively mitigate numerous problems, such as “dimensional disaster” and traditional pattern recognition (Farber et al., 2016; Li, Ni, et al., 2018; Nieto et al., 2013; Qingyang et al., 2012; Sun et al., 2011, 2014; Tao et al., 2020; Yang et al., 2017). The general linear regression model can be expressed as follows:

\[
f(x) = w^T x + b
\]

where \( w \) is the normal vector of the API input vector, and \( b \) is the deviation value. The loss value is zero only when \( f(x) \) is exactly the same as the true value. As we know, in the actual air passenger index forecast, it is impossible to predict the exact value of each day accurately. However, the SVR model “softens” the prediction result, allowing a certain error between the predicted and actual values, which is equivalent to forming a prediction error isolation band with a width of \( 2\varepsilon \) at the center of the prediction value, falling into the isolation band. If the API value prediction is accurate, the loss is 0, and the API input vector closest to the isolation zone constitutes its “support vector.” Noteworthy is that minimizing the loss requires maximizing the sum \( \rho \) of the distances between the two sets of support vectors and the prediction center, which can be achieved by minimizing the Euclidean norm of the normal vector \( w \). Thus, the SVR concept can be expressed as
where $C$ is a regularization constant for performing a compromise calculation on the front and the back. The former term (front) indicates that all predicted values fall within the error range, as much as possible in the model structure. The latter (back) applies the $\varepsilon$-insensitive loss function to characterize the fit between the model prediction effect and the actual passenger volume data.

$$l_i(z) = \begin{cases} 0 & \text{if } |z| \leq \varepsilon \\ |z| - \varepsilon & \text{otherwise} \end{cases}$$ (10)

In the actual API data, a certain value may exceed the normal trend due to external reasons and become an outlier. In this case, the “hard interval” defined above is no longer applicable. Therefore, in the case of serious deviations from the actual value, slack variables $\xi_i$ and $\xi_i^*$ are introduced as “softening” intervals, which reduces the problem formulation to the following form:

$$\min \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{m} (\xi_i - \xi_i^*)$$

s.t. \[
\begin{align*}
    f(x_i) - y_i &\leq \varepsilon + \xi_i \\
    y_i - f(x_i) &\leq \varepsilon + \xi_i^* \\
    i &\in \{1, 2, \ldots, m\}; \xi_i, \xi_i^* \geq 0
\end{align*}
\] (11)

Using the dual principle and introducing Lagrangian multipliers $\alpha_i$ and $\alpha_i^*$, the SVR’s dual problem can be formulated as in [38]:

$$\max_{\alpha, \alpha^*} \sum_{i=1}^{m} y_i (\alpha_i^* - \alpha_i) - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} y_i y_j (\alpha_i^* + \alpha_j) x_i^T x_j$$

s.t. \[
\sum_{i=1}^{m} (\alpha_i^* - \alpha_i) = 0, 0 \leq \alpha_i^*, \alpha_i \leq C
\] (12)

When the predicted value of the API falls into the $\varepsilon$-soft zone, $\alpha_i$ and $\alpha_i^*$ can be the non-zero value. Insofar as the predicted value cannot simultaneously fall into two opposite areas, at least one of the parameters $\alpha_i$ or $\alpha_i^*$ is 0. Finally, the SVR regression prediction function (Zhou, 2016) can be expressed as:

$$f(x) = \sum_{i=1}^{m} (\alpha_i^* - \alpha_i) x_i^T x + b$$

$$b = y_i - \varepsilon - \sum_{j=1}^{m} (\alpha_j - \alpha_j^*) x_j^T x_i$$ (14)

For the API time series data with a nonlinear trend, the SVR can map the sample to the high-dimensional space through the nonlinear mapping function $\phi(x)$, and then replace the inner vector product of the high-dimensional space $\phi(x_i) \cdot \phi(x_j)$ with the kernel function $K(x_i, x_j)$. The most commonly used kernel function is the Gaussian radial basis kernel function (RBF) (Zhou, 2016), which can be expressed as follows:

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

where gamma is the Gaussian radial basis kernel function parameter (gamma=$\frac{1}{2\sigma^2}$) and $\sigma > 0$ is the Gaussian kernel bandwidth.

The RBF function application improves the SVR nonlinear prediction ability. Eventually, the SVR regression function takes the following form (Zhou, 2016):

$$f(x) = \sum_{i=1}^{m} (\alpha_i^* - \alpha_i) K(x, x_i) + b$$ (16)

**Model Elaboration**

Based on the preliminary data preprocessing, the raw data were converted into the corresponding information entropy values. All data were then normalized to permit the elimination of dimensions in the data unrelated to the API. These dimensions/factors were defined as those with small influence according to the ranking of MI values. Next, the key influencing factors were selected to set the ML model’s foundation based on the MI principle and improve the prediction given smaller dimensions. The graphical representation of the MI-SVR model elaboration procedure is given in Figure 1.

The following steps were required for constructing the MI-SVR model for the API prediction.

**Step1:** According to the API’s definition in section 2, airport passenger throughput data were processed to get the original sequence $\{X_i\}$.

**Step2:** The influencing factors of API were transformed into information entropy, and then the standardized information entropy was obtained by standardizing (normalizing) the fixed unit length of the converted data.

**Step3:** According to the mathematical definition of API in section 2, the airport passenger throughput data were converted into the API sequence as $\{X_i^*\}$.

**Step4:** The K-means method was used to cluster API and influencing factors (maximum temperature, minimum temperature, wind force, and wind direction) when forming different clusters.

**Step5:** The standardized information entropy of each influencing factor, and its MI value with LAPI value was calculated.
Figure 1. MI-SVR model elaboration process and its application to API prediction: (a) the key influencing factors selected based on mutual information and (b) blended MI-SVR model application process.
Step 6: According to the characteristics of MI value and the correlation, the factors with high MI values were selected. These factors were introduced into the elaborated MI-SVR model to predict the API.

**Empirical Analysis**

**Numerical Experiments**

Data acquisition and processing. Using the API definition in Section 2, the raw data was considered representative and suitable for modeling without missing values. The data used for the experiments had to be converted into the form consistent with the constructed model’s input dimension. In this study, the original air passenger throughput values were converted to API. The dataset used in this paper included the complete raw data on daily air passenger throughput, maximum temperature, minimum temperature, weather, wind direction, and wind power for the PVG, which covered 20 months from January 1, 2017, and August 31, 2018. The weather, wind direction, and wind power data had a textual format, while the temperature data were graded to indicate different temperature levels. The data conversion results are summarized in Table 1.

For the airport operation data modeling based on the theory of information entropy, this study converted the five categories of data into their corresponding information entropy values. Then, the total passenger throughput data were calculated according to formula (1), and the standardized passenger volume value was classified as the API by the K-means clustering algorithm. The classification results are listed in Table 2. Finally, the MI values of the transfer passenger travel index level and the five key influencing factors were calculated separately. The latter factors controlling the API were the maximum temperature, minimum temperature, wind direction, weather, and wind power. The MI values of these factors in the API are given in Table 3.

From the Table 3, it can be concluded that the least impact on the API is wind power. Through statistical analysis of wind factors, it was revealed that the proportion of wind level above level 5 is only 2.3%, while nearly 88% of the data is not higher than level 3 (level 3–4). Therefore, the influence impact of wind power on the nearly API is relatively small.

Error analysis method. To analyze different model prediction results, this study used the mean absolute percentage error (MAPE) and root mean square error (RMSE), which can be derived via the following equations:

\[
\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - y_i^*}{y_i} \right| \times 100\% \tag{17}
\]

\[
\text{RMSE} = \sqrt{\frac{1}{n} \sum_{j=1}^{n} (y_j - y_j^*)^2} \tag{18}
\]

where \( y_i \) and \( y_i^* \) are the actual and predicted values, respectively. In the experiment, the actual \( y_i \) and predicted \( y_i^* \) API
values of different models are calculated, and MAPE and RMSE are calculated to evaluate the superiority of different models for API prediction.

**Discussion**

**Selection of key factors.** The RMSE and MAPE values obtained via the MI-SVR model were 0.1030 and 11.44% for the maximum temperature as a single influencing factor. Those for the minimum temperature were 0.1016 and 11.18%, respectively. The minimum temperature effect was found to be slightly higher than that of the maximum one. The analysis also showed that the Pearson correlation coefficient between the maximum and minimum temperatures was 0.581, and a close correlation within the 99% confidence interval was observed.

From Table 3, it can be concluded that wind power had the least impact on the API. The wind power factor’s statistical analysis revealed that the share of recorded wind power above level 5 was only 2.3%, while the remaining 88% of the data corresponded to levels 3 and 4. Therefore, the wind power effect on the API was relatively weak.

Given these findings, three key influencing factors, namely the minimum temperature, weather, and wind direction, were selected for incorporation into the forecast of the air passenger index to get more accurate prediction results.

The experimental data used in this study were subdivided into two parts. The first part covered 19 months (577 days) of API data from January 1, 2017, to July 31, 2018, and was applied to the model training. The second part covered the remaining month (August 1–31, 2018) of the dataset and was adopted as a test set for verifying the model fitting effect. The numerical experiments were realized via the Python 3.6 software and produced some noteworthy results. After numerous numerical experiments, it is found that when the specific parameters of the model were set at certain values, the error terms of the overall effect of both the training and test sets reached their minimum values. The model-related parameter settings are listed in Table 4.

**Comparative experimental analysis.** This study proposed blended MI-SVR model was tested with the account of influencing factors and without their account (as a single original SVR model). For comparison, commonly used LSTM and ARIMA single model, as well as their blended variants, namely the GM(1, 1)-BPNN and SVR-ARIMA models (Liang et al., 2015; Tuo, 2012), were applied to the training and test datasets. The actual and predicted values of the above six prediction models were calculated, as shown in Figure 2, where the left part of the black dotted line depicts the respective model fitting effect on the training set, and the right part corresponds to the prediction results on the test set. The MI-SVR model outperformed the LSTM and ARIMA single model, which only met the general trend of the data but did not fit well with the higher and lower values. The blended MI-SVR model had a better prediction effect than single models by introducing the conditional entropy of influencing factors.

The MAPE and RMSE values of the six models calculated via equations (17) and (18), respectively, are listed in Tables 5 and 6. It is seen that all blended models had a better prediction effect than single ones and others blended variants models, such as GM(1, 1)-BPNN and SVR-ARIMA models.

In summary, since MI is the random event correlation analysis method, the MI value can track the existence of a potential relationship. Thus, screening key influencing factors for API prediction based on MI provided the necessary conditions for improving the model’s predictive accuracy. After the analysis of each model and the introduction of the conditional entropy of influencing factors, the overall prediction effect of each model was significantly improved. It is noteworthy that the MI-SVR model designed in this paper had the best prediction accuracy among other tested single or blended prediction models.

**Air traffic volume analysis.** The developed MI-SVR model was applied to analyze the API evolution in the PVG during a more extended period, namely 130 months, from January 2008 to October 2018. For brevity’s sake, only the most prominent features, which are considered instrumental in guiding operational and managerial decisions, are presented in this study. Figure 3 depicts the LAPI values in each month.

- January and February of each year correspond to the “dead season” of the PVG, and API values are the lowest in those months, which closely correlates with the Chinese New Year festivities, during this time, many people prefer to reunite with their families rather than travel. Airport operation managers can refer to the LAPI to make optimal operation and maintenance plans. In other words, this period can be used by the airport management for performing expansion or upgrade works on its core and ancillary facilities.
- Fairly stable API values are observed in the period from March to June, November, and December.
months are relatively busy months for major airlines, this period airport can arrange fixed service resources to guarantee operations, and airlines can draw up scientific flight plans.

✓ The busiest period for each year is from July to October. From a global standpoint, this period corresponds to the summer vacation travel period. From a local standpoint, this period is the most lucrative for Shanghai’s major tourist attractions.

Conclusions and Future Research Avenues

Conclusions
The current study attempts to deal with the operational and managerial challenges induced by the airport passenger throughput increase. In order to achieve this goal, the data on airport passenger throughput were converted into the
corresponding API and LAPI values. They were evaluated by the K-means clustering method, which indicates the passenger flow at civil aviation airports. The LAPI provides certain decision-making references for the general public to choose travel time and transportation options.

Aiming at the prediction of API in civil aviation airports, this paper proposes a blended MI-SVR model, takes account of the key influencing factors for improving the prediction results. Influencing factors selected in this study includes minimum temperature, weather conditions, and wind direction. By way of experimental simulation, the model was verified on the daily airport passenger throughput data of PVG. The prediction results of the proposed MI-SVR model were compared with those of such popular ML prediction models, such as LSTM, ARIMA, SVR-ARIMA, and GM(1,1)-BPNN models. Based on the proposed MI-SVR method, experiments were carried out on historical airport passenger throughput. Experimental results illustrate the effectiveness and advantages of the proposed method. The main findings of the comparison results are as follows:

1. The results provided by the proposed model outperform other tested single or blended models by the overall prediction accuracy.
2. The original SVR for predict problems is extended, and the MI-based influencing factors of API prediction are scientifically viable.
3. In contrast to single (LSTM and ARIMA) models and blended (SVR-ARIMA and GM(1,1)-BPNN) models, the proposed MI-SVR model achieved relatively better prediction results.
4. The LAPI evolution at PVG conforms to the rising annual population patterns.

| Table 5. RMSE and MAPE Comparison of Different Models. |
|----------------|------|--------|
| Model          | RMSE | MAPE (%) |
| SVR + influencing factors (MI-SVR) | 0.0785 | 8.04 |
| LSTM + influencing factors | 0.1053 | 11.86 |
| ARIMA + influencing factors | 0.1060 | 11.66 |
| SVR             | 0.1031 | 11.33 |
| LSTM            | 0.1064 | 12.10 |
| ARIMA           | 0.1093 | 12.09 |

| Table 6. Comparison With the Prediction Effect of Others Blended Variants Model. |
|----------------|------|--------|
| Model          | RMSE | MPAE (%) |
| SVR-ARIMA      | 0.0981 | 10.41% |
| GM(1,1)-BPNN   | 0.1032 | 11.45% |
| MI-SVR         | 0.0785 | 8.04% |

Figure 3. The LAPI of PVG from January 2008 to October 2018.
Finally, the MI-SVR model provides an effective reference method for evaluating API in airport construction, operation, and management. For example, the maintenance schedule of airport facilities and equipment, ground service, and aircraft allocation can be optimized with API. The effectiveness of MI-SVR model is demonstrated for API prediction and the results are very satisfactory. This study will relieve both airport operators and managers, as well as passengers, of possible performance deterioration.

**Future Research Avenues**

In the follow-up studies, more possible influencing factors data would be collected and processed to establish a dynamic air trip prediction model. Besides, more datasets, such as high-speed trains network, highway network, air traffic control, and origin-destination demand, will be incorporated into the methodological framework. Furthermore, we plan to investigate the applicability of this methodology to datasets of other civil airports, in other words, to improve the model training quality, there is a need for more data. We also encourage other researchers to explore these directions for civil aviation development.

**Acknowledgments**

We thank the editors and any reviewers for their helpful comments.

**Author Contributions**

Conceptualization and Formal analysis, Honglin Xiong; Writing—discussion of the original draft, Chongjun Fan and Collins Opoku Antwi; Data curation, Yun Yang; Methodology, Chongjun Fan; redrafting and editing, Collins Opoku Antwi and Xiaomao Fan; discussion of reviewer(s)’ Comments and giving important suggestions, Chongjun Fan and Hongmin Chen.

**Declaration of Conflicting Interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by Key (Key grant) Project of Chinese Ministry of Education (Grant No. 20JZD010), the National Natural Science Foundation of China (Grant No. 171774111), and the Key fund projects of Shanghai Municipal Education Commission for scientific research and innovation (Grant No. 14ZZ131).

**ORCID iD**

Honglin Xiong https://orcid.org/0000-0002-7714-100X

**References**

Barth, M., Emrich, E., & Güllich, A. (2019). A machine learning approach to “revisit” specialization and sampling in institutionalized practice. *Sage Open, 9*(2), 2158244019840554.

Başaran, S., & Ejimogu, O. H. (2021). A neural network approach for predicting personality from Facebook data. *Sage Open, 11*(3), 21582440211032156.

Cai, K., Li, Y., Fang, Y. P., & Zhu, Y. (2021). A deep learning approach for flight delay prediction through time-evolving graphs. *IEEE Transactions on Intelligent Transportation Systems, 8*, 1–11.

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access, 8*, 75264–75278.

Chernyshov, K. R. (2013). An information theoretic approach to constructing machine learning criteria. *IFAC Proceedings Volumes, 46*(11), 269–274.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.

Dantas, T. M., Cyrino Oliveira, F. L., & Varela Repolho, H. M. (2017). Air transportation demand forecast through Bagging Holt Winters methods. *Journal of Air Transport Management, 59*, 116–123.

Fan, X. L., Feng, H. H., & Yuan, M. (2013). Principal component analysis feature selection algorithm based on mutual information. *Control and Decision, 28*(6), 915–919.

Farber, S., Ritter, B., & Fu, L. (2016). Space–time mismatch between transit service and observed travel patterns in the Wasatch Front, Utah: A social equity perspective. *Travel Behaviour and Society, 4*, 40–48.

Gao, L., & Wu, W. (2020). Relevance assignment feature selection method based on mutual information for machine learning. *Knowledge-Based Systems, 209*, 106439.

Greff, K., Srivastava, R. K., Koutnik, J., Steunebrink, B. R., & Schmidhuber, J. (2017). LSTM: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems, 28*(10), 2222–2232.

Gunning, D., & Aha, D. (2019). DARPA’s explainable artificial intelligence (XAI) program. *AI Magazine, 40*(2), 44–58.

Guo, S., & Zhang, G. (2020). Using machine learning for analyzing sentiment orientations toward eight countries. *Sage Open, 10*(3), 2158244020951268.

Jun, L., XueFeng, Y., YingHua, L., ZhangZhi, C., ZheJun, K., XiaoCai, H., & Di, D. (2010). Thermo-viscoelastic analysis of the integrated T-shaped composite structures. *Composites Science and Technology, 70*(10), 1497–1503.

Kinney, J. B., & Atwal, G. S. (2014). Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences, 111*(9), 3354–3359.

Kong, W., Dong, Z. Y., Jia, Y., Hill, D. J., Xu, Y., & Zhang, Y. (2017). Short-term residential load forecasting based on LSTM recurrent neural network. *IEEE Transactions on Smart Grid, 10*(1), 841–851.

Liang, C. Y., Ma, Y. C., & Chen, R. (2015). Prediction of daily tourism demand based on SVR-ARIMA combination model. *Journal of Management in Engineering, 30*(1), 122–127.

Li, P., Ni, Z. W., & Zhu, X. H. (2018). Air pollution index prediction model of support vector machine based on fractal manifold learning. *Systems Science and Mathematics, 38*(11), 1296–1306.

Li, Q. (2014). Grey Verhulst model in commercial flights at Macau international airport. *Journal of Grey System, 26*(2), 170.

Li, Y., & Cao, H. (2018). Prediction for tourism flow based on LSTM neural network. *Procedia Computer Science, 129*, 277–283.

Li, Y. H., Han, H. Y., Liu, X., & Li, C. (2018, September 21–23). *Passenger flow forecast of Sanya airport based on ARIMA...*
model [Conference session]. International Conference of Pioneering Computer Scientists, Engineers and Educators, Zhengzhou, China (pp. 442–454). Springer.

Li, Y., & Jiang, X. (2020). Airport passenger throughput forecast based on PSO-SVR model. *IOP Conference Series: Materials Science and Engineering*, 780(6), 062006.

Ma, C. W., & Ma, Y. G. (2018). Shannon information entropy in heavy-ion collisions. *Progress in Particle and Nuclear Physics*, 99, 120–158.

Nieto, P. G., Combarro, E. F., del Coz Díaz, J. J., & Montañés, E. (2013). A SVM-based regression model to study the air quality at local scale in Oviedo urban area (northern Spain): A case study. *Applied Mathematics and Computation*, 219(17), 8923–8937.

Niu, B., Ren, J., Zhao, A., & Li, X. (2020). Lender trust on the P2P lending: Analysis based on sentiment analysis of comment text. *Sustainability*, 12, 3293.

Qingyang, C., Zhenping, S., Daxue, L., Yuqiang, F., & Xiaohui, L. (2012). Local path planning for an unmanned ground vehicle based on SVM. *International Journal of Advanced Robotic Systems*, 9(6), 246.

Sharmin, S., Shoyaib, M., Ali, A. A., Khan, M. A. H., & Chae, O. (2019). Simultaneous feature selection and discretization based on mutual information. *Pattern Recognition*, 91, 162–174.

Sulistyowati, R., Kuswanto, H., & Astuti, E. T. (2018, March 6–7). Hybrid forecasting model to predict air passenger and cargo in Indonesia [Conference session]. 2018 International Conference on Information and Communications Technology (ICOIAC), Yogyakarta, Indonesia (pp. 442–447). IEEE.

Sun, H., Yang, P. R., & Cheng, J. H. (2011). Energy demand prediction model based on Matlab support vector regression machine. *Systems Engineering - Theory & Practice*, 31(10), 2001–2007.

Sun, S., Lu, H., Tsui, K. L., & Wang, S. (2019). Nonlinear vector auto-regression neural network for forecasting air passenger flow. *Journal of Air Transport Management*, 78, 54–62.

Sun, S., Wei, Y., Tsui, K. L., & Wang, S. (2019). Forecasting tourist arrivals with machine learning and internet search index. *Tourism Management*, 70, 1–10.

Sun, Y. X., Shao, C. F., & Ji, X. (2014). Time series prediction of traffic accidents based on ARIMA and information granulation SVR combination model. *Journal of Tsinghua University (Natural Science Edition)*, 3, 348–353.

Tao, D., Ma, Q., Li, S., Xie, Z., Lin, D., & Li, S. (2020). Support vector regression for the relationships between ground motion parameters and macroseismic intensity in the Sichuan–Yunnan region. *Applied Sciences*, 10(9), 3086.

Tuo, Q. (2012). Application of combination model in airport passenger throughput prediction. *Computer Simulation*, 29(4), 108–111.

Ullal, M. S., Havaldar, I. T., Soni, R., & Nadeem, M. (2021). The role of machine learning in digital marketing. *Sage Open*, 11(4), 21582440211050394.

Xia, L., Jie, Y., Lei, C., & Ming-Rui, C. (2016, October 13–15). Prediction for air route passenger flow based on a grey prediction model [Conference session]. 2016 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CYBERC), Chengdu, China (pp. 185–190). IEEE.

Yang, H. F., Dillon, T. S., & Chen, Y. P. (2017). Optimized structure of the traffic flow forecasting model with a deep learning approach. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10), 2371–2381.

Zhang, C. T., Ma, Q. L., & Peng, H. (2016). Prediction of chaotic time series based on information entropy optimization for phase space reconstruction parameters. *Journal of Physics*, 59, 7623–7629.

Zhou, Z. H. (2016). *Machine learning* (pp. 121–135). Tsinghua University Press.

Zou, Y., Zhang, G., & Liu, L. (2019). Research on image steganography analysis based on deep learning. *Journal of Visual Communication and Image Representation*, 60, 266–275.