An integrated approach to air passenger index prediction: mutual information principle and support vector regression (MI-SVR) blended model

Honglin Xiong [1], Chongjun Fan[1, *], Collins Opoku ANTWI[1], Yun Yang[1] and Xiaomao Fan[2]

[1] Business School, University of Shanghai for Science and Technology, Shanghai 200093, China; honyex@126.com(H.X.); scantwi28@yahoo.com (C.A.); 2424586590@qq.com(Y.Y.)

[2] School of Computer Science, South China Normal University, Guangzhou 510631, China; xmfan@scnu.edu.cn(X.F.)

* Correspondence: chongjunfan@126.com (C.F.); Tel.: +86-180-17060-389

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. Based on the 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 passenger throughput data of the Shanghai Pudong International Airport, China. The experimental results prove the model feasibility and effectiveness by comparing them with conventional methods, such as ARIMA, LSTM, and other machine learning models, outperformed by the MI-SVR model. Besides, it is shown that the prediction effect of each model could be improved by introducing influencing factors based on mutual information. 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(ML); mutual information(MI); support vector regression (SVR); K-Means

1. Introduction

The global 'digital divide" status quo is quickly changing with the progress in artificial intelligence (AI) technologies and their application area expansion. The heart of AI technologies is machine learning (ML), which has branched into shallow and deep learning [1]. Examples of shallow ML models are the support vector machine (SVM) invented by Cortes and Vapnik [2] in 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 [3-7]. 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 [8-10]. Such linear models as autoregressive integrated moving average (ARIMA) [10] and gray model (GM) [11] have achieved significant predictive results. However, nonlinear models, including the BP-neural network [12], recurrent neural network (RNN) [13], long- and short-term memory network (LSTM) [14-16], and support vector regression (SVR) [17-21] 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 [9,40], 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 Shanghai Pudong International Airport (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 [22, 23]. 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.

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 [24], while classical neural network models and other blended models are exploited for the hybrid prediction of airport passenger and freight route flows [25]. The regional aviation passenger flow aspect is covered by the neural networks and SVR models [9], 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 [21]. Air passenger flow characteristics are quite accurately predicted based on the time sequence columns [5,15,16,19] and nonlinear vector autoregressive neural network [9]. In study [26]. the 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 [22, 27] and ML field [23, 39-40] is well-known. In contrast to the previous studies, this paper analyzes the passenger throughput of a civil aviation airport and processes the 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 mutual information to build an ML prediction model, to find a better prediction method for the travel schedule selection by 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 theory and then presents the blended MI-SVR model. Section 4 analyzes the passenger throughput and weather data of the Shanghai Pudong International Airport (PVG), China, to verify the MI-SVR model's feasibility and effectiveness and assess the PVG air passenger index. Section 5 concludes the paper and discusses future research avenues.

2. Research Tops

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 passenger throughput to the air passenger index (API), which takes into account the airport's capacity and

95 variability in passenger traffic volumes. The following definitions provide the mathematical
96 representation of API and the level of air passenger index (LAPI).

97 **Air passenger index (API):** Setting the value of as the airport passenger throughput per unit
98 period, the API for this period is defined as X_t^*:

$$X_t^* = \frac{X_t - X_{min}}{X_{max} - X_{min}} \tag{1}$$

100 where X_{min} and X_{max} are the minimum and maximum numbers of air passengers per unit time,
101 respectively, while X_t^* varies between 0 and 1.

102 **Level of air passenger index (LAPI):**Set $\{p_1, p_2, p_3 \cdots p_t\}$ is the API sequence set of multiple
103 time units. After clustering, the generated cluster $\{N_t\}$ is a collection of data objects. The value is
104 the data element of each cluster after clustering. If the API's unit period is one month, and the
105 airport's monthly API is X_t^*, the level of air passenger index (LAPI) is derived as:

$$N_{X_t^*} = \begin{cases} 1, p_t \in (0, i) \\ 2, p_t \in (i, j) \\ 3, p_t \in (j, k) \\ \dots \\ N, p_t \in (\theta, 1) \end{cases} \tag{2}$$

107 Where p_1, p_2, p_t are time units, while i, j, k, θ are cluster boundary values.

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

3. Related Theory and Model Elaboration

3.1. Mutual Information Theory

113 The information entropy is a key indicator in the information theory, which was introduced by
114 Shannon in 1948[28] based on the following concept: the more ordered is a system, the smaller is its
115 information entropy and vice versa. Therefore, information entropy can be used to measure the
116 degree of system uncertainty (or degree of ordering) [28-31]. Information entropy can be derived via
117 the following formula:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i) \tag{3}$$

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

122 Let the joint probability distribution of the random vector (X, Y) be p_{ij}, then the
123 two-dimensional joint entropy of vector (X, Y) is:

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

125 Assuming that the joint probability distributions of X and Y are $p_{i\cdot}$ and $p_{\cdot j}$, respectively, the
126 conditional entropy can be defined as:

$$H(X / Y) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} \log \frac{p_{ij}}{p_{\cdot j}}$$

$$H(Y / X) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} \log \frac{p_{ij}}{p_{i\cdot}} \tag{5}$$

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

130
$$\begin{aligned} I(X;Y) &= H(X) - H(X \mid Y) \\ &= H(Y) - H(Y \mid X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned} \tag{6}$$

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

133
$$I(X;Y) = \sum_{i,j} p_{ij} \log_2 \frac{p_{ij}}{p_{i\bullet} p_{\bullet j}} \tag{7}$$

134   *3.2. Support Vector Regression (SVR)*

135   The support vector regression (SVR) is a widely used ML prediction method, which adopts the
136   principle of minimizing structural risk rather than minimizing empirical risk. This allows one to
137   effectively mitigate numerous problems, such as "dimensional disaster" and traditional pattern
138   recognition [17-19,32-36]. The general linear regression model can be expressed as follows：

139
$$f(x) = w^T x + b \tag{8}$$

140   where w is the normal vector of the API input vector, and b is the deviation value. The loss
141   value is zero only when $f(x)$ is exactly the same as the true value. In this study, In the actual air
142   passenger index forecast, it is impossible to predict the exact value of each day accurately. However,
143   the SVR model "softens" the prediction result, allowing a certain error between the predicted and
144   actual values, which is equivalent to forming a prediction error isolation band with a width of $2* \varepsilon$
145   at the center of the prediction value, falling into the isolation band. If the API value prediction is
146   accurate, the loss is 0, and the API input vector closest to the isolation zone constitutes its "support
147   vector." Noteworthy is that minimizing the loss requires maximizing the sum r of the distances
148   between the two sets of support vectors and the prediction center, which can be achieved by
149   minimizing the Euclidean norm of the normal vector w. Thus, the SVR concept can be expressed as

150
$$\min \frac{1}{2} w^2 + C \sum_{i=1}^m l_\epsilon \left(f(x_i) - y_i \right), \quad C > 0 \tag{9}$$

151   where C is a regularization constant for performing a compromise calculation on the front and the
152   back. The former term (front) indicates that all predicted values fall within the error range, as much
153   as possible in the model structure. The latter (back) applies the ε-insensitive loss function to
154   characterize the fit between the model prediction effect and the actual passenger volume data.

155
$$l_\epsilon(z) = \begin{cases} 0 & \text{if } |z| \le \varepsilon \\ |z| - \varepsilon & \text{otherwise} \end{cases} \tag{10}$$

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

160
$$\min \frac{1}{2} \| w \|^2 + C \sum_{i=1}^m (\xi_i - \xi_i^*) \tag{11}$$

161
$$s.t. \begin{cases} f(x_i) - y_i \le \varepsilon + \xi_i \\ y_i - f(x_i) \le \varepsilon + \xi_i^* \end{cases} \quad i = 1, 2, \ldots, m; \xi_i, \xi_i^* \ge 0$$

162   Using the dual principle and introducing Lagrangian multipliers α_i and α_i^*, the SVR's dual
163   problem can be formulated as in [37]:

164
$$\max_{\alpha, \alpha^*} \sum_{i=1}^m y_i (\alpha_i^* - \alpha_i) - \varepsilon(\alpha_i^* + \alpha_i) - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m y_i (\alpha_i^* - \alpha_i)(\alpha_j^* + \alpha_j) x_i^T x_j \tag{12}$$

165
$$s.t.\sum_{i=1}^{m}(\alpha_i^* - \alpha_i) = 0, 0 \le \alpha_i^*, \alpha_i \le C$$

166 When the predicted value of the API falls into the ε-soft zone, α_i and α_i^* can be the non-zero
167 value. Insofar as the predicted value cannot simultaneously fall into two opposite areas, at least one
168 of the parameters α_i or α_i^* is 0. Finally, the SVR regression prediction function [37] can be expressed
169 as:

170
$$f(x) = \sum_{i=1}^{m}(\alpha_i^* - \alpha_i)x_i^T x + b \tag{13}$$

171
$$b = y_i + \in - \sum_{j=1}^{m}(\alpha_j - \alpha_j^*)x_j^T x_i \tag{14}$$

172 For the API time series data with a nonlinear trend, the SVR can map the sample to the
173 high-dimensional space through the nonlinear mapping function $\varphi(x)$, and then replace the inner
174 vector product of the high-dimensional space $\varphi(x_i) \cdot \varphi(x_j)$ with the kernel function $K(x_i, x_j)$. The
175 most commonly used kernel function is the Gaussian radial basis kernel function(RBF) [37], which
176 can be expressed as follows:

177
$$K(x_i, x_j) = \exp(-\frac{\| x_i - x_j \|^2}{2\sigma^2}) \tag{15}$$

178 where *gamma* is the Gaussian radial basis kernel function parameter (*gamma*=$\frac{1}{2\sigma^2}$) and
179 $\sigma > 0$ is the Gaussian kernel bandwidth.
180 The RBF function application improves the SVR nonlinear prediction ability. Eventually, the SVR regression
181 function takes the following form [37]:

182
$$f(x) = \sum_{i=1}^{m}(\alpha_i^* - \alpha_i)K(x, x_i) + b \tag{16}$$

3.3. Model Elaboration

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

191
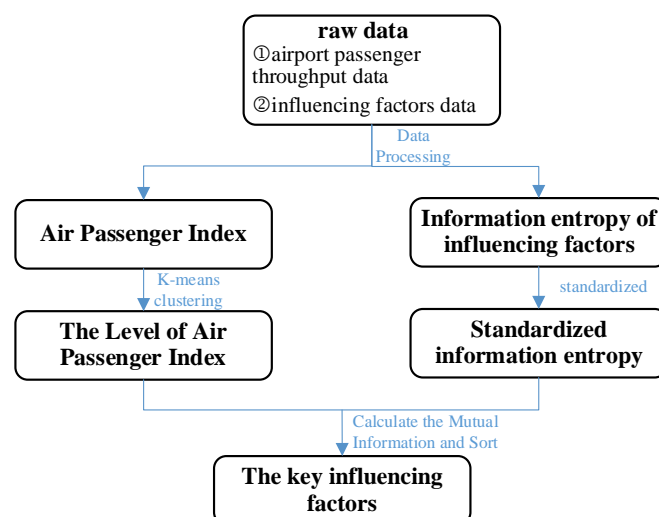
192   (a) The key influencing factors selected based on mutual information

193
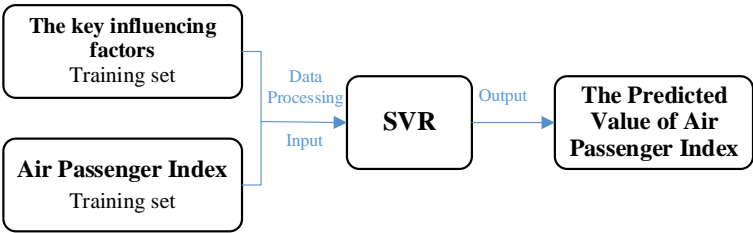
194   (b) Blended MI-SVR model application process

195   **Figure 1.** MI-SVR model elaboration process and its application to API prediction

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

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

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

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

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

207   **Step5:** The standardized information entropy of each influencing factor, and its MI value with
208   LAPI was calculated.

209   **Step6:** According to the characteristics of MI value and the correlation, the factors with high MI
210   values were selected. These factors were introduced into the elaborated MI-SVR model to predict the
211   API.

212   **4. Empirical Analysis**

213   *4.1. Numerical Experiments*

214   4.1.1. Data Acquisition and Processing

215   Using the API definition in Section 2, the raw data was considered representative and suitable
216   for modeling without missing values. First, the data had to be converted into the form consistent
217   with the constructed model's input dimension. In our case, the original passenger throughput values
218   were converted to API. The dataset used in this paper included the complete raw data on daily
219   passenger throughput, maximum temperature, minimum temperature, weather, wind direction,
220   and wind power for the Shanghai Pudong International Airport (PVG), China, which covered 20
221   months from January 1, 2017, and August 31, 2018. The weather, wind direction, and wind power
222   data had a textual format, while the temperature data were graded to indicate different temperature
223   levels. The data conversion results are summarized in Table1.

224   **Table 1.** Data preprocessing correspondence table and pre-processing raw data from this table

Maximum temperature	Minimum temperature	Wind power	Weather	Wind direction
2-10（cold）	-3-4（Very cold）	Level 1-2	Sunny/Cloudy	No sustained wind direction
11-17（Micro cold）	5-10（cold）	Level 2-3	Cloudy	East Wind
18-25（Moderate）	11-17（Micro cold）	Level 3-4	Light Rain	South Wind

26-32（Micro heat）	18-24（Moderate）	Level 4-5	Shower	West Wind
33-40（heat）	25-31（Micro heat）	Level 5 or higher	Rain	North Wind
			Heavy Rain	Southeast Wind
			Rainstorm	Northeast Wind
				Southwest Wind
				Northwest Wind

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

233   **Table 2.** Using K-means clustering algorithm to classify API , and the API is finally divided into 5 levels.

Level	Travel Index Range	Description
1	0-0.13	Smooth
2	0.13-0.35	Less Smooth
3	0.35-0.52	A little Congestion
4	0.52-0.79	Congestion
5	0.79-1	Severe Congestion

234   **Table 3.** MI between influencing factor and the API, and the ranking of the MI values.

Influencing factor	Mutual information value with air passenger index	Rank
Maximum temperature	0.616	1
Minimum temperature	0.534	2
Weather	0.249	4
Wind direction	0.290	3
Wind power	0.136	5

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

239   4.1.2. Error Analysis Method

240   To analyze the model prediction results, we used the mean absolute percentage error (MAPE)
241 and root mean square error (RMSE), which can be derived via the following equations:

242
$$\text{MAPE} = \frac{\sum_{i=1}^{n} \frac{\left| y_i - y_i^* \right|}{y_i}}{n} \times 100\% \tag{17}$$

243
$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - y_i^*)^2} \tag{18}$$

244   where y_i and y_i^* are the actual and predicted values, respectively.

245   *4.2. Discussion*

246   4.2.1. Selection of key factors

247   The RMSE and MAPE values obtained via the MI-SVR model were 0.1030 and 11.44% for the
248   maximum temperature as a single influencing factor. Those for the minimum temperature were
249   0.1016 and 11.18%, respectively. The minimum temperature effect was found to be slightly higher
250   than that of the maximum one. The analysis also showed that the Pearson correlation coefficient
251   between the maximum and minimum temperatures was 0.581, and a close correlation within the
252   99% confidence interval was observed.
253   From Table 3, it can be concluded that wind power had the least impact on the API. The wind
254   power factor's statistical analysis revealed that the share of recorded wind power above level 5 was
255   only 2.3%, while the remaining 88% of the data corresponded to levels 3 and 4. Therefore, the wind
256   power effect on the API was relatively weak.
257   Given these findings, three key influencing factors, namely the minimum temperature,
258   weather, and wind direction, were selected for incorporation into the forecast of the airline
259   passenger travel index to get more accurate prediction results.
260   The experimental data used in this study were subdivided into two parts. The first part covered
261   19 months (577 days) of API data from January 1, 2017, to July 31, 2018, and was applied to the
262   model training. The second part covered the remaining month (August 1-31, 2018) of the dataset and
263   was adopted as a test set for verifying the model fitting effect. The numerical experiments were
264   realized via the Python 3.6 software and produced some noteworthy results. After numerous
265   numerical experiments, it is found that when the specific parameters of the model were set at certain
266   values, the error terms of the overall effect of both the training and test sets reached their minimum
267   values. The model-related parameter settings are listed in Table 4.

268   **Table 4.** SVR model parameter setting.

Kernel function	Gamma in kernel function	Penalty parameter C	ε
RBF	0.64	6000	0.1

269   4.2.2. Comparative Experimental Analysis

270   The proposed blended MI-SVR model was tested with the account of influencing factors and
271   without their account (as a single SVR model). For comparison, commonly used LSTM and ARIMA
272   single model, as well as their blended variants, namely the GM(1, 1)-BPNN and SVR-ARIMA
273   models [20, 21], were applied to the training and test datasets. The actual and predicted values of the
274   above six prediction models were calculated, as shown in Figure 2, where the left part of the black
275   dotted line depicts the respective model fitting effect on the training set, and the right part
276   corresponds to the prediction results on the test set. The MI-SVR model outperformed the LSTM and
277   ARIMA single models, which only met the general trend of the data but did not fit well with the
278   higher and lower values. The blended MI-SVR model had a better prediction effect than single
279   models by introducing the conditional entropy of influencing factors.
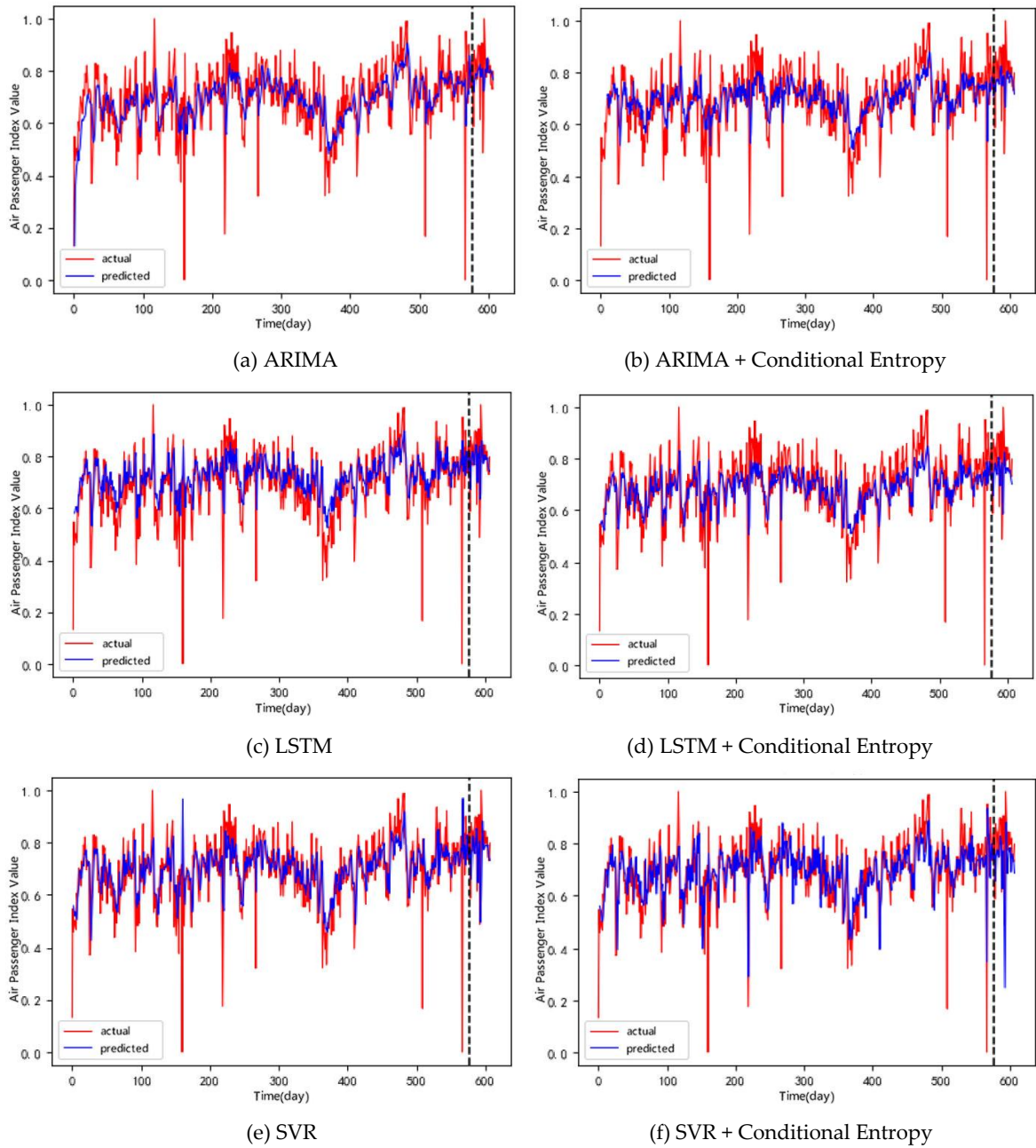
280
281   (a) ARIMA   (b) ARIMA + Conditional Entropy

282
283   (c) LSTM   (d) LSTM + Conditional Entropy

284
285   (e) SVR   (f) SVR + Conditional Entropy

286   **Figure 2.** Comparison of actual and predicted values of the different model

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

290

291   **Table 5.** RMSE and MPAE 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%

In summary, since MI is the random event correlation analysis method, the mutual information 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 highest temperature, the lowest temperature, and the wind direction as 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.

4.2.3. Air Traffic Volume Analysis

The developed MI-SVG model was applied to analyze the API evolution in the Shanghai Pudong International Airport (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.
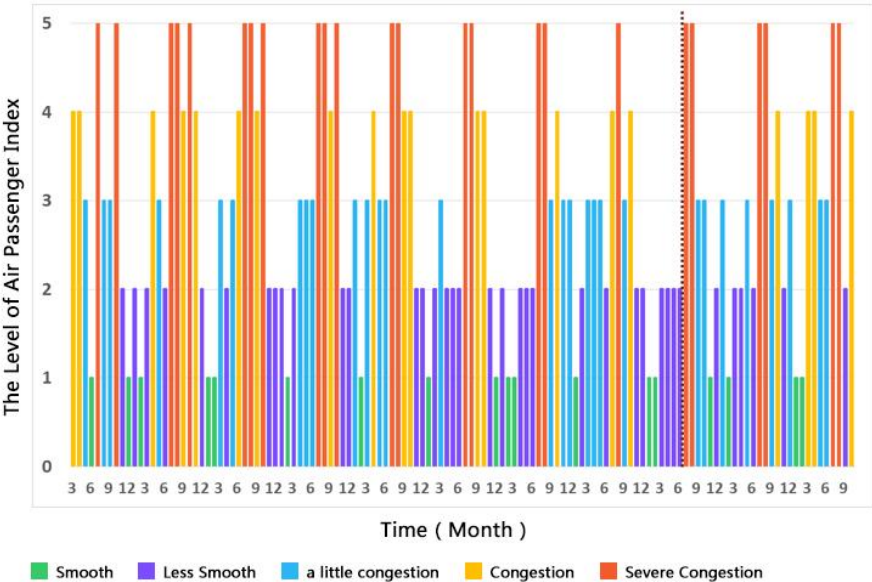

Figure 3. The Level of Air Passenger Index (LAPI) of PVG from January 2008 to October 2018sssss

- ✓ January and February of each year correspond to the "dead season" of the PVG airport, and their API values are the lowest, which closely correlates with the Chinese New Year festivities. 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. These months are relatively busy months for major airlines.
- ✓ 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

316   period is the most lucrative for Shanghai's major tourist attractions.

317   **5. Conclusions and Future Research Avenues**

318   *5.1. Conclusions*

319   The current study attempts to deal with the operational and managerial challenges induced by
320   the airport passenger throughput increase. To achieve this objective, the data on airport passenger
321   throughput were converted into the corresponding API and LAPI values. They were evaluated by
322   the K-means clustering method, which indicates the passenger flow at civil aviation airports. The
323   LAPI provides certain decision-making references for the general public to choose travel time and
324   transportation options.
325   Aiming at the prediction of API in civil aviation airports, this paper proposes a blended MI-SVR
326   model, takes account of the key influencing factors for improving the prediction results. Influencing
327   factors selected in this study includes minimum temperature, weather conditions, and wind
328   direction. By way of experimental simulation, the model was verified on the daily passenger
329   throughput data of PVG. The prediction results of the proposed MI-SVR model were compared with
330   those of such popular ML prediction models, such as LSTM, ARIMA, SVR-ARIMA, and
331   GM(1,1)-BPNN models. The main findings of the comparison results are as follows:
332   (1) The results provided by the proposed model outperform other tested single or blended
333   models by the overall prediction accuracy;
334   (2) The MI-based influencing factors of API prediction are scientifically viable;
335   (3) In contrast to single (LSTM and ARIMA) models and blended (SVR-ARIMA and
336   GM(1,1)-BPNN) models, the proposed MI-SVR model achieved relatively better prediction results;
337   (4) The LAPI evolution at PVG conforms to the rising annual population patterns.
338   Finally, the MI-SVR model provides an effective reference method for evaluating API in airport
339   construction, operation, and management. For example, the maintenance schedule of airport
340   facilities and equipment, ground service, and aircraft allocation can be optimized with API. This will
341   relieve both airport operators and managers, as well as passengers, of possible performance
342   deterioration.

343   *5.2. Future Research Avenues*

344   In the follow-up studies, more data would be collected and processed to establish a dynamic air
345   trip prediction model. Besides, more datasets, such as high-speed trains network, highway network,
346   air traffic control, and origin-destination demand, will be incorporated into the methodological
347   framework. Furthermore, we plan to investigate the applicability of this methodology to datasets of
348   other civil airports. We also encourage other researchers to explore these directions for civil aviation
349   development.

350   **Author Contributions:** Conceptualization and Formal analysis, H.X.; Writing—discussion of the original draft
351   ,H.X. and H.J.; Data curation, Y.Y.; Methodology, C.F.; redrafting and editing, C.A. and X.F..

356   **Conflicts of Interest:** The authors declare no conflict of interest.

357   **References**

358   1.   Manuel Woschank.; Erwin Rauch.; Helmut Zsifkovits. Machine Learning, and Deep Learning in Smart
359   Logistics. *Sustainability* **2020**, *12*, 3760.
360   2.   Cortes C.; Vapnik V. Support-vector networks. *Machine Learning*. **1995**, *20*, 273-297.
361   3.   Schultz, M.; Schmidt, M. Advancements in Passenger Processes at Airports from Aircraft Perspective.
362   *Sustainability* **2018**, *10*, 3877.

363 4. Y. Zou.; G. Zhang.; L. A. Liu. Research on image steganography analysis based on deep learning. *Journal of*
364   *Visual Communication and Image Representation*. **2019**, *60*, 266-275.

365 5. Hyojoo Son.; Changwan Kim. A Deep Learning Approach to Forecasting Monthly Demand for
366   Residential–Sector Electricity. *Sustainability*. **2020**, *12*, 3103.

367 6. Chen, L. J.; Chen, P. P.; Z. J. Lin. Artificial intelligence in education: a review. *IEEE Access*. **2020**, *8*,
368   75264-75278.

369 7. Hyojoo Son.; Changwan Kim. A Deep Learning Approach to Forecasting Monthly Demand for
370   Residential–Sector Electricity. *Sustainability* **2020**, *12*, 3103.

371 8. X. J. Feng.; Wei X.; Y. L. Huang. Predicting the Airport Passenger Throughput Based on a Support Vector
372   Regression Mode. *Computer Engineering*. **2005**, *31*, 172-173.

373 9. S. L. Sun.; H. X. Lu.; Kwok-Leung Tsui, et al. Nonlinear vector auto-regression neural network for
374   forecasting air passenger flow. *Journal of Air Transport Management.*, **2019**, *78*, 54-62.

375 10. L. Zhang.; S. F. Yan. Comparison of Holt- Winters and ARIMA Methods for Forecasting Charge of China
376   Airline Passengers. *Shanghai University of Engineering Science*. **2006**, 20, 280-283,.

377 11. Yao Y B.; Gao H. Gray model G M(l, 2) Application in airport passenger throughput pre-side. *Journal of*
378   *Civil Aviation Flight University of China*. **2006**, *17*, 12-16.

379 12. Jiao P P. Forecasting Method and Its Mechanism of Impacts on Airport Passenger Throughput.
380   *Transportation Systems Engineering and Information Technology*. **2005**, *1*, 107-110.

381 13. Fatma Yaprakdal.; M. Berkay Yılmaz.; Mustafa Baysal .;Amjad Anvari-Moghaddam. A Deep Neural
382   Network-Assisted Approach to Enhance Short-Term Optimal Operational Scheduling of a Microgrid.
383   *Sustainability* **2020**, *12*, 1653.

384 14. Klaus Greff.; Rupesh Kumar Srivastava.; Jan Koutnik; et al. LSTM: A Search space odyssey. *IEEE*
385   *Transactions on Neural Networks and Learning Systems*. **2017**, *28*, 2222–2232.

386 15. W. Kong.; Z. Y. Dong.; Jia Y. Dong, *et al*. Short-Term Residential Load Forecasting based on LSTM
387   Recurrent Neural Network. *IEEE Transactions on Smart Grid*. **2019**, *10*, 841- 851.

388 16. F.L. Yi.; C. Han. Prediction for Tourism Flow based on LSTM Neural Network. *Procedia Computer Science*.
389   **2018**, *129*, 277-283.

390 17. P. Li.; Z. W. Ni.; X. H. Zhu; *et al*. Air pollution index prediction model of support vector machine based on
391   Fractal manifold learning. *Systems Science and Mathematics*. **2018**, *38*, 1296-1306.

392 18. H. Sun.; P. R. Yang.; J. H. Cheng. Energy demand prediction model based on Matlab support vector
393   regression machine. *Systems engineering theory and practice*. **2011**, *31*, 2001-2007.

394 19. Y. X. Sun.; C. F. Shao.; X. Ji; *et al*. Time series prediction of traffic accidents based on ARIMA and
395   information granulation SVR combination model. *Journal of Tsinghua University (natural science edition)*.
396   **2014**, *3*, 348-353.

397 20. C. Y. Liang.; Y. C. Ma.; R. Chen; *et al*. Prediction of daily tourism demand based on SVR- ARIMA
398   combination model. *Journal of management engineering*. **2015**, *29*, 122-127.

399 21. Qu, T. Application of combination model in airport passenger throughput prediction. *Computer*
400   *simulation.*, **2012**, *4*, 108-111.

401 22. Sadia Sharmin.; Mohammad Shoyaib.; Amin AhsanAli, *et al*. Simultaneous feature selection and
402   discretization based on mutual information. *Pattern Recognition*. **2019**, *91*, 162-174.

403 23. K.R.Chernyshov. An Information Theoretic Approach to Constructing Machine Learning Criteria. *IFAC*
404   *Proceedings Volumes*. **2013**, *46*, 269-274.

405 24. Xia, L.; Jie, Y.; C. Lei; *et al*. Prediction for Air Route Passenger Flow Based on a Grey Prediction Model.
406   *International Conference on Cyber-enabled Distributed Computing & Knowledge Discovery IEEE*, Chengdu,
407   China, **2016**.

408 25. Ratna Sulistyowati.; Suhartono.; Heri Kuswanto.; *et al*. Hybrid forecasting model to Predict air passenger
409   and cargo in Indonesia. *International Conference on Information and Communications Technology (ICOIACT)*.
410   *IEEE*., Zhengzhou, China, **2018**, pp.442-447.

411 26. Li, J.; Yao, X. F.; Liu Y H, *et al*. Thermo-viscoelastic Analysis of the Integrated T-shaped Composite
412   Structures. *Composites Science and Technology*. **2010**, *70*, 1497-1503.

413 27. C E Shannon. A Mathematical theory of communication. *The Bell System Technical Journal*. **1948**, pp.
414   379-423.

415 28. Fan, X. L.; Feng, H. H.; Ma, Y. Principal Component Analysis Feature Selection Algorithms Based on
416   Mutual Information. *Control and Decision-making.*, **2013**, *28*, 915-919.

417   29.   Liang, J. Y.; Feng, C. J.; Song, P. Summary of Big Data Relevance Analysis. *Journal of Computer Science.*,
418   **2016**, *1*, 1-18.
419   30.   Wang, Y. Research on Machine Learning Method Based on Mutual Information and Prior Information.
420   Ph.D. dissertation, Graduate School of Chinese Academy of Sciences (Institute of Automation), Beijing,
421   China, **2008**.
422   31.   Zhang, C. T.; Ma, Q. L.; Peng, H. Prediction of chaotic time series based on information entropy
423   optimization for phase space reconstruction parameters. *Journal of Physics.* **2016**, *59*, 7623-7629.
424   32.   Tao, D. W.; Ma, Q.; Li, S. L.; Xie, Zh. N.; Lin, D. X. *et al.* Support Vector Regression for the Relationships
425   between Ground Motion Parameters and Macro seismic Intensity in the Sichuan–Yunnan Region. *Appl.*
426   *Sci.* **2020**, *10*, 3086.
427   33.   Xu, Y. Research on prediction algorithm of abnormal trend of spacecraft in-orbit state based on machine
428   learning method" M.S. thesis, University of Electronic Science and Technology, Chengdu, China, **2017**.
429   34.   Steven Farber.; Benjamin Ritter.; Fu, L.W. Space–time mismatch between transit service and observed
430   travel patterns in the Wasatch Front, Utah: A social equity perspective. *Travel Behaviour and Society.* **2016**, *4*,
431   40-48.
432   35.   Zhou, W.; Wang, W.; Hua, X.D. ; Zhang, Y. Real-Time Traffic Flow Forecasting via a Novel Method
433   Combining Periodic-Trend Decomposition. *Sustainability* **2020**, *12*, 5891.
434   36.   J. García Nieto.; E.F. Combarro.; J.J. Del Coz Díaz, *et al.* A SVM-based regression model to study the air
435   quality at local scale in Oviedo urban area (Northern Spain): A case study. *Applied Mathematics &*
436   *Computation.* **2013**, *219*, 8923-8937.
437   37.   Zhou, Z. H. *Machine Learning*, Tsinghua University Press, China: Beijing, **2016**, pp. 121-135.
438   38.   Biljana Petrevska. Predicting tourism demand by ARIMA models. *Economic Research-Ekonomska*
439   *Istrazivanja.* **2017**, *30*, 939-950.
440   39.   Wu, W. W.; Zhang, H. Y.; Feng, T.; Frank, Witlox. A Network Modelling Approach to Flight Delay
441   Propagation: Some Empirical Evidence from China. *Sustainability* **2019**, *11*, 4408.
442   40.   Sun, S. L.; Wei, Y. J.; Kwok-Leung Tsui. *et al.* Forecasting tourist arrivals with machine learning and
443   internet search index. *Tourism Management.* **2019**, *70*, 1-10.
444