1. \documentclass[12pt]{article}
2. % \usepackage{setspace} % For line spacing if needed
3. \input{settings/mycommands.tex}
4. \input{settings/packages.tex}
5. \input{settings/settings.sty}
6. \title{ Comparative Study of Time Series Models for Temperature Forecasting in Delhi }
7. \author{Feng Gu(T00751197), Yishu Liu(T00728937), Haoran He(T00749480)}
8. % \date{\today}
9. \begin{document}
10. \maketitle
11. \begin{center}
12. \textbf{**\large Abstract**}
13. \end{center}
14. **This study investigates regional temperature forecasting in Delhi, using climate data collected from 1st January 2013 to 24th April 2017. Various time series models, including dynamic regression model and linear regression, with and without dummy variables, alongside benchmark models like naive, drift, and mean forecasts were applied to the data. We evaluated the models’ performance using metrics such as RMSE, MAE, and MAPE. Results indicate that the dynamic regression model with SARIMA errors and dummy variables outperforms other models, achieving the lowest RMSE (3.0829) and MAPE (12.4737). These findings highlight the effectiveness of incorporating dummy variables in improving temperature prediction accuracy, offering insights for future applications in climate data modeling and decision-making.**
15. }
16. **\section{Introduction}**
17. \sloppy
18. The subject of global warming and climate change is gradually becoming one of the significant challenges
19. that the world must face. More frequent and intense extreme weather events, such as heat waves, dust storms,
20. and floods, have been observed globally \cite{dabhade2021}.
21. The issue of climate change has become particularly crucial in large, densely populated cities.
22. One such example is Delhi, India. The effects of climate change have intensified in recent years,
23. posing challenges to human health, agricultural production, and the environment
24. \cite{hussain2024}.
25. Therefore, studying the temperature trends in Delhi holds high scientific value and practical significance.
26. \sloppy
27. This study employs time series analysis, enabling the examination of historical temperature data,
28. comparison of different time series models, and the prediction of future temperature trends using the
29. best-fit model. This research aims to provide a comprehensive understanding of temperature trends in the Delhi region,
30. and to explore and practice a more efficient way to organize and finish the paper writing work.
31. **\section{Data}**
32. **\subsection{Source of data}**
33. The climate data for the city of Delhi, India, spanning from 1st January 2013 to 24th April 2017, was downloaded from Kaggle\footnote{a machine learning community for learners} and originally sourced from Weather Undergroud API,
34. The dataset consists 1576 records with date index and other 4 variables: mean temperature,
35. humidity, wind speed and mean pressure. The mean temperature is the target variable and
36. the other variables are used as predictors.
37. **\subsection{Preparing and processing the data}**
38. We process the raw data by following the procedure below.:
39. \begin{itemize}
40. \item We check the missing values in the training data and fill them using linear interpolation.
41. \item We explore the distribution of the 4 variables, with boxplots and histogram shown in Figure 2 and 3. Additionally, STL decomposition is applied to analyze the trend, seasonality and remainder of data, providing better understanding for the data.
42. \item The abnormal outliers\footnote{outliers were detected using customized algorithm, which could be
43. found in the code of Module}
44. are replaced with corresponding moving average values.
45. \item We create dummy variables from the "date" variable: four seasons.
46. \item Before the model fitting, we perform the stationary check and determine that the training data requires first-order differencing.
47. \end{itemize}
48. **\section{Method}**
49. The complete code, latex documents and images can be found in the following \textbf{**GitHub repo:**}
50. \href{https://github.com/Gufeng-2002/Final-report-for-time-series.git}
51. {https://github.com/Gufeng-2002/Final-report-for-time-series.git}
52. **\subsection{Specifing the desired model}**
53. Before we set a specific model for forecasting \textit{*meantemp*},
54. we decomposed the \textit{*meantemp*} using TSL method\cite{fpp3stl}. Becasue
55. we have daily climate data, we set the season period as 365, assuming
56. the same day in each year should have the most similar pattern in Temperature
57. \footnote{But it is not rigirous, because every
58. four-year there is one more day premium and the number of day
59. is not an accurate "365" of interge.}.
60. After observing the possible seasonality and trend, we create a assume the model as following:
61. \[
62. y\_t = \beta\_{5 \times 1} X + \beta\_{3 \times 1} H + \eta\_t
63. \]
64. in which:
65. \[
66. X = \begin{bmatrix}
67. 1 & 1 & X\_{11} & X\_{21} & X\_{31}\\
68. 1 & 2 & X\_{12} & X\_{22} & X\_{32}\\
69. & & ... & ... &\\
70. 1 & t & X\_{1t} & X\_{2t} & X\_{3t} \\
71. \end{bmatrix}
72. \quad H\_i =
73. \begin{bmatrix}
74. 1 & 0 & 0\\
75. 0 & 1 & 0\\
76. & ... & \\
77. 0 & 0 & 0 \\
78. \end{bmatrix} =
79. \begin{cases}
80. 1, \text{if the season is i} \\
81. 0, other wise
82. \end{cases}
83. \]
84. There are totally three $H\_i$ here to aviod multilinearity
85. caused by including intercept. To the $\eta\_t$, we assume it follows
86. a SARIMA or ARIMA model, specificly:
87. \[
88. \Phi^P(B^s) \phi^p(B) (1-B^s)^D (1-B)^d \eta\_t =
89. \Theta^Q(B^s) \theta^q(B) \epsilon\_t
90. \]
91. where, we set the $s$ equal to 365(days). The searching for peroper order
92. of SARIMA((P,D,Q) and (p,d,q)) and the specific claculating are finished
93. by R language.
94. **\subsection{Comparisions with other models}**
95. In order to assess our model properly, we totally build \textbf{**eight**} models: Mean, Drift,
96. Naive, Saive, Linear model with dummy variables or not, dynamic regression model with dummy variables or not
97. , shown in table 3, appendix.
98. **\subsection{Complete workflow}**
99. \subsubsection{ProcessRawData module of Python}
100. It is notable that the data processing steps are finished in a workflow with module \textit{*"ProcessRawData.py"*}
101. \cite{financialriskforecasting},
102. which has been pushed to the public Git repository. It can be easy
103. \footnote{only needing to point or change the directory path correctly}
104. to repeat all these steps or make
105. further adjustments to make it suitable for other work.
106. \subsubsection{ModelFitting module of R}
107. To fit these models quickly and easily, we choose R to build these models and
108. do relevant tests on them and visualize the results.
109. There is a \textit{*"ModlFitting.R"*} in the repo. There are
110. some functions that transport tables from R to Latex document, which
111. accelerated our work.
112. **\section{Results}**
113. The specific settings about parameters of models can be found in the R module.
114. According the table 4 and 7, we compared the performance of these models on
115. training data and testing data, the dynamic regression model with dummy variables performs
116. well on training and testing data sets, its \textbf{**RMSE(3.0829)**} and \textbf{**MAPE(12.4737)**} are
117. the lowest in all models(standard linear regression with dummy variables of \textbf{**RMSE(3.6619)**} and \textbf{**MAPE(13.7899)**}).
118. The AICc and log\\_lik are higher than linear models', but it is mainly becasue the
119. number of parameters is more than models', which is reasonable.
120. Information about this model is shown in table 1.
121. From table 4 and 5 in appendix, we found that models with dummy variabes are always
122. better in performance than models without that.
123. The four varialbes: \textit{*time, season\\_Autumn, season\\_Spring, season\\_Summer*}
124. pass the 0.05 significance level test under the
125. $H\_{0}\footnote{$H\_0$: the coefficient is value of 0, namely no influcence from this variable}$
126. assumption, however, they do not pass the corresponding tests in dynamic regression models.
127. To the reason why these variables become not important in dynamic regression,
128. one explanation might be that the influences from these four variables can be captured
129. well by the errors of SARIMA process in the model, and the long-term trend with \textit{*time*} is also
130. not imporant to mean temperature, based on the given sample.
131. Additionally, there is one counterintuitive coefficent: the coefficient
132. for \textit{*season\\_Summer*} is smaller than that of \textit{*season\\_Spring*}, which should
133. not be correct by checking the summary about the average feature value in table 1(appendix).
134. It might indicate that some predictors in our model take the effect from \textit{*season\\_Summer*},
135. if we remove these interfering factors, the relationship might be shown correctly, or this is
136. the truth of the real word.
137. Although some coefficients did not pass the significance level test, we can still
138. use the model to forecast, becasue we are focusing on the relationship between these
139. variables but the future values of target.
140. \begin{table}[!h]
141. \centering
142. \captionsetup{font=small} % Set caption to left-align and smaller font
143. \caption{\textit{*Summay about the dynamic regression model.*
144. *Including the coefficents, tests about residuals from training data,*
145. *and criteria about performance from testing data.*
146. *(Note: the dynamic regression model here is called 'sarima\\_dummy' in R code*
147. *and tables in appendix)*}}
148. \label{tab:model\_summary\_combined}
149. \begin{tabular}{lccccc}
150. \toprule
151. \textbf{**Metric**} & \textbf{**ME**} & \textbf{**RMSE**} & \textbf{**MAE**} & \textbf{**MPE**} \\
152. \midrule
153. \multirow{2}{\*}{dynamic regression}
154. & 0.5447 & 3.0829 & 2.6184 & -0.079 \\
155. \cmidrule{2-5}
156. & \textbf{**MAPE**} & \textbf{**ACF1**} & \textbf{**log\\_lik**} & \textbf{**AIC**} \\
157. \cmidrule{2-5}
158. & 12.4737 & 0.8543 & -2369.349 & 4762.699 \\
159. \midrule
160. \textbf{**Coefficient**} & \textbf{**Estimate**} & \textbf{**Std. Error**} & \textbf{**Statistic**} & \textbf{**P-value**} \\
161. \cmidrule{1-5}
162. ar1 & 0.9898 & 0.0041 & 242.1087 & 0.0000 \\
163. ma1 & -0.0953 & 0.0298 & -3.2015 & 0.0014 \\
164. ma2 & -0.1798 & 0.0300 & -5.9982 & 0.0000 \\
165. humidity & -0.1363 & 0.0042 & -32.4098 & 0.0000 \\
166. wind\\_speed & -0.0291 & 0.0072 & -4.0637 & 0.0001 \\
167. meanpressure & -0.0322 & 0.0076 & -4.2461 & 0.0000 \\
168. time & 0.0021 & 0.0045 & 0.4730 & 0.6363 \\
169. season\\_Autumn & 0.2608 & 0.5227 & 0.4990 & 0.6179 \\
170. season\\_Spring & 0.5930 & 0.5235 & 1.1326 & 0.2576 \\
171. season\\_Summer & 0.4116 & 0.6098 & 0.6751 & 0.4997 \\
172. intercept & 63.9278 & 8.5701 & 7.4594 & 0.0000 \\
173. \midrule
174. \textbf{**Other Metrics**} & \textbf{**sigma2**} & \textbf{**log\\_lik**} & \textbf{**AICc**} & \textbf{**BIC**} \\
175. \cmidrule{1-5}
176. \multirow{2}{\*}{dynamic regression} & 1.5048 & -2369.349 & 4762.914 & 4826.149 \\
177. \cmidrule{2-5}
178. & \textbf{**lb\\_stat**} & \textbf{**lb\\_pvalue**} & \textbf{**bp\\_stat**} & \textbf{**bp\\_pvalue**} \\
179. \cmidrule{2-5}
180. & 1.5524 & 0.2128 & 1.5492 & 0.2133 \\
181. \bottomrule
182. \end{tabular}
183. \end{table}
184. We also visualized the forecasts for all models to
185. make the comparisons more clear and direct in figure 1.
186. \begin{figure}[!h]
187. \centering
188. \includegraphics[width=.8\textwidth]{images/forecasts\_CI90.png}
189. \captionsetup{font=small} % Set caption to left-align and smaller font
190. \caption{\textit{*Forecasts from eight models.*
191. *Becasue of the assumptions and settings to models,*
192. *we should compare the closes forecasts from dynamic regression model with forecasts*
193. *from other models.*}}
194. \label{fig:figure1}
195. \end{figure}
196. **\section{Discussion}**
197. **\subsection{Explanation about the model results}**
198. According the regression results, we found that \textit{*humidity, wind speed and mean pressure*} have
199. negative effect on mean temperature, with their increases, the temperature decreases.
200. In comparison with the winter, the other seasons have higher mean temperature,
201. even thought the coefficent of \textit{*Spring*} and \textit{*Summer*} might look counterintuitive,
202. which could be the task for further exploration.
203. The autoregression and moving average parts show there are strong autocorrelation
204. in the mean temperature variable, which could be explained by standard linear model
205. that considers \textit{*time and seasons*} variables in some extent.
206. **\subsection{Other useful work and further improvement.}**
207. We have to admit it is not a very rigirous report due to the lack of time and
208. the limits of our skills and professional knowledge in coding and Time Series field.
209. However, this report is a try in using Vscode, Rstudio, Latex entention as a complete
210. workflow, in which we manage to finish all the work in one system and make the
211. whole process automatic as much as possible. The complete frame work could be found in the GitHub
212. repository, including the document frame of Latex.
213. To make the whole workflow better, i think we can make improvement with the following
214. aspects:
215. \begin{itemize}
216. \item Learn Time Series forecasting models more
217. \item Be familiar with R and Python for Data Science
218. \item Be familiar with using Vscode, Rstudio and GitHub for collaboration.
219. \end{itemize}
220. \clearpage
221. % Bibliography section
222. \begin{thebibliography}{99} % The number specifies the width of the label
223. \bibitem{dabhade2021}
224. A. Dabhade, S. Roy, M. S. Moustafa, S. A. Mohamed, R. El Gendy, and S. Barma,
225. ``Extreme Weather Event (Cyclone) Detection in India Using Advanced Deep Learning Techniques,''
226. \textit{*2021 9th International Conference on Orange Technology (ICOT)*},
227. Tainan, Taiwan, 2021, pp. 1--4,
228. doi: \href{https://doi.org/10.1109/ICOT54518.2021.9680663}{10.1109/ICOT54518.2021.9680663}.
229. \bibitem{hussain2024}
230. Hussain S., Hussain E., Saxena P., Sharma A., Thathola P., Sonwani S.,
231. ``Navigating the impact of climate change in India: a perspective on climate action (SDG13) and sustainable cities and communities (SDG11),''
232. \textit{*Frontiers in Sustainable Cities*}, 2024 Jan 23.
233. Available from:
234. \href{https://research-ebsco-com.ezproxy.tru.ca/linkprocessor/plink?id=6c6991da-78bd-3062-a586-5a5d83ba7467}{https://research-ebsco-com.ezproxy.tru.ca}.
235. \bibitem{financialriskforecasting}
236. A. McNeil.
237. ``Financial Risk Forecasting: R Best Practice,''
238. \textit{*Financial Risk Forecasting Notebook*}.
239. Available at: \url{https://www.financialriskforecasting.com/notebook/R/BestPractice.html}.
240. Accessed: November 30, 2024.
241. \bibitem{fpp3stl}
242. H. Hyndman and G. Athanasopoulos.
243. ``STL Decomposition,''
244. \textit{*Forecasting: Principles and Practice (3rd ed.)*}.
245. Available at: \url{https://otexts.com/fpp3/stl.html}.
246. Accessed: November 30, 2024.
247. \end{thebibliography}
249. % \input{sections/appendix\_figure.tex}
250. \input{sections/appendix\_table.tex}
251. \end{document}