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