STAT450/550 Multivariate Statistical Analysis

Term Project

Note:

1. Late projects will NOT be accepted.
2. You have to write the project paper as if you are submitting a statistical report to the company you are consulting, assuming that they are not statisticians. Your paper must be concise, precise, and easy to read (please type in a Word Process).
3. You may choose to use your own data of interest. If this is the case, you are asked to submit a brief description of the data and objectives of the study (one or two pages) by April 14.
4. Carefully labeled and captioned outputs (e.g., Table 1, .., Figure 1,…) should be inside of your paper or attach as appendices. The SAS outputs must be well summarized and organized in tables and figures. Do not turn in the raw SAS outputs. Resize figures and tables. Do NOT turn in any outputs not discussed.
5. To receive a full credit you must include complete discussions and your paper must be brief and well organized. Do not exceed 10 pages, excluding title page, tables and figures.

Data

The South Coast Air Basin, which covers Los Angeles, Orange, Riverside, and San Bernardino counties in Southern California, currently is not in compliance with National Ambient Air Quality Standard (NAAQS) for ozone. According to the American Lung Association, the four counties graded “F” for ozone and San Bernardino ranked as the most ozone-polluted county in the U.S. Riverside and Los Angeles ranked the third and the fourth, respectively. The California Air Resources Board (ARB) gathers air quality data for the State of California. In 1987 California ARB approved the daily maximum of 90 ppb standard. It is well known that significant harmful health effects could occur among both adults and children if exposed to levels above the standards.

Data are based on daily maximum ozone records (in ppb) from five monitoring stations in San Bernardino County from January 2001 to December 2004. The stations are listed in Table 2. It is well known that the formation of ozone is heavily dependent on meteorological conditions. The meteorological data came from the University of California Statewide Integrated Pest Management Program (UC IPM: www.ipm.ucdavis.edu). It consists of daily ambient and soil temperature, global radiation, relative humidity, evapotranspiration, and wind speed and direction (see Table 3). Note that data from only two stations (SB, Upland) are posted. Each data separated by two periods (2001~2003 and 2004). Use the first three-years data for model fitting and the 2004 data for the purpose of validating your model.

US Environmental Protection Agency (US EPA) provides daily Air Quality Index (AQI) to public to rate air quality of given area (http://www.airnow.gov/). Currently, the AQI is classified to six categories: Good, Moderate, Unhealthy for sensitive groups, Unhealthy, Very unhealthy, and Hazardous depending on the level of public health concern. AQI forecasts, if reliable and accurate, could play an important role as part of a local air quality management system. However, statistical prediction of AQI is currently not available.

The *purpose* of this study is to implement multivariate statistical methods to classify daily AQI based on the meteorological conditions. The AQI index with the corresponding ranges of Ozone concentration is given in Table1. Using the table, convert the daily ozone records to AQI states. At the minimum, your project paper should include a preliminary analysis via covariance or correlation matrix, a complete PCA or FA for the meteorological variables (Table 3, variable 1~8), a complete DA and Classification. You may also consider Logistic regression approach for the classification. The use of month and weekend variables is your choice. If you use month, you need to convert it to 6 dummy variables, or ozone season (1 if month is May through September, 0 otherwise) .

Table 1. Summary table of AQI state by US EPA

Levels of Health Concern	AQI State	Ozone ranges (ppb)
Good	1	0-59
Moderate	2	60-75
Unhealthy for Sensitive Group	3	76-95
Unhealthy	4	96-115
Very unhealthy	5	116-374
Hazardous	6	375+

Table 2: Ozone monitoring stations in San Bernardino County, CA

Stations	AIRS ID	Latitude	Longitude	Missing (01~04)	%
Crestline	060710005	34.24	-117.28	2%	
Fontana	060712002	34.10	-117.49	4%	
Redlands	060714003	34.06	-117.15	0%	
SB- 4th St.	060719004	34.11	-117.27	1%	
Upland	060711004	34.10	-117.63	0%	

Table 3: Meterological variables and others

1. temp.max:	Daily maximum air temperature (F)
2. eto:	Daily reference evapotranspiration (inch)
3. prec:	Daily total precipitation (inch)
4. rh.max:	Daily maximum relative humidity (%)
5. soil.max:	Daily maximum soil temperature (F)
6. solar:	Daily global radiation (Watts/m^2)
7. wind.u:	Daily west-east wind component (m/s)
8. wind.v:	Daily south-north wind component (m/s)
9. month	1=Jan, 2=Feb, etc
10. weekend	0=weekday, 1=weekend (Saturday, Sunday)