Journal of the Korean Data & Information Science Society
2015, **26**(2), 399–408

http://dx.doi.org/10.7465/jkdi.2015.26.2.399
한국데이터정보과학회지

R 소프트웨어를 이용한 대기오염 데이터의 시각화[†]

오영창[1] · 박은식[2]

[12]전남대학교 통계학과

접수 2015년 2월 3일, 수정 2015년 2월 12일, 게재확정 2015년 3월 19일

요 약

본 논문은 대기오염 자료를 여러 가지 방법의 데이터 시각화를 통해 나타내었고, 데이터 시각화 방법별로 통계적인 방법을 활용한 분석과 연계하여 어떤 특징을 알아 볼 수 있는지를 나타냈다. 데이터 시각화 도구로는 통계 패키지인 R을 사용하였다. 분석에 사용된 데이터는 뉴욕시에서 1973년 5월부터 9월까지 공기의 질을 측정한 자료이다. 먼저 단변량 분석과 단순회귀분석을 실시하여 데이터 시각화를 통해 자료의 기본적인 특성을 파악하고 시각화 방법으로 산점도행렬 등을 통해 특성을 한눈에 볼 수 있게 나타내었다. 다중 회귀 분석을 실시하여 로그변환 등을 이용하여 최적의 모형을 찾고 설명변수들을 범주화하여 상자그림이나 3차원 투시도, 3차원 산점도 등 여러 데이터 시각화 방법을 이용해 대기오염 데이터의 전체적인 특성들을 알아보았다.

주요용어: 데이터 시각화, 산점도 행렬, 상자그림, 3차원 산점도, 3차원 투시도.

1. 서론

빅데이터 시대를 맞이하여 이에 관한 조사 및 연구가 끊임없이 진행되고 있고 국가와 기업에서는 빅데이터 시스템을 구축하고 활용하기 위한 많은 노력이 진행 중에 있다 (Lim 등, 2012). 빅데이터 시장은 시간이 지날수록 더욱 더 커지고 있으며 (Bae 등, 2013), 일상생활에서도 다양한 분야에서 이용되고 일반 사람들에게도 많은 정보가 공유되고 있다. 하지만 빅데이터 분석은 매우 복잡하고 난해하여 자료가 의미 하는 것이 무엇인지 어떤 방향성을 나타내는지 이해하기 어려운 부분이 많기 때문에 빅데이터의 시각화가 중요해지고 있다 (Choi 등, 2013).

데이터의 특징을 쉽게 파악하기 위한 효율적인 수단으로 시각화를 들 수 있다 (Cho, 2014; Park, 2014). 데이터가 어떠한 특징을 갖고 있는지 쉽게 이해하기 힘든 일반인들도 한눈에 볼 수 있는 데이터 시각화가 자료의 이해를 쉽게 해준다. 다양한 시각화 기법들이 개발되고 있으며 (Joo 등, 2013), 여러 가지 컴퓨터 프로그램들을 사용하여 구현 할 수 있다. 일반인들도 무료로 접근이 가능하고, 최신 통계분석도 다양하며 그래픽 성능이 우수한 R 프로그램은 데이터 시각화에 적합한 도구라고 할 수 있다 (Kim과 Lee, 2014). R 프로그램을 통해 다양한 방법으로 데이터를 시각화 하여 자료를 한눈에 볼 수 있도록 하고 그에 따른 정보를 쉽게 알 수 있게 하여 효과적으로 데이터를 파악하는데 도움이 되고자 한다 (Park, 2007).

데이터 시각화에 이용된 자료는 R 프로그램에 내재되어 있는 “Airquality”라는 자료이다 (Chambers 등, 1983). 이 자료를 이용하여 오존의 농도에 영향을 주는 기상 자료들과의 연관성을 분석하고자 한다.

[†] 이 논문은 2014년도 전남대학교 학술연구비 지원에 의하여 연구되었음.
[1] (500-757) 광주광역시 북구 용봉로 77, 전남대학교 통계학과 석사과정.
[2] 교신저자: (500-757) 광주광역시 북구 용봉로 77, 전남대학교 통계학과, 교수. E-mail: espark02@jnu.ac.kr

데이터를 파악하기 위하여 각 변수별로 단변량 분석을 실시하고 각 변수 간의 관계를 살피기 위해서 단순 회귀 분석과 다중 회귀 분석을 실시하여 변수 간의 관계를 파악한다. R 프로그램을 이용하여 단변량 자료의 그래프 및 삼변량 이상의 2차원 및 3차원 그래프를 구현하여 다중 회귀 분석 결과와 비교해 보고자 한다.

이 글의 2절에서는 "Airquality" 자료에 대한 자세한 설명이 되어 있다. 3절에서는 자료를 통계적으로 분석하였고 4절에서는 분석 결과를 쉽게 이해할 수 있는 데이터 시각화 방법들로 나타내었다. 마지막 5절에서는 분석 후 결론을 제시하였다.

2. 자료 설명

"Airquality"데이터는 R 프로그램에 내재되어 있는 자료로써 1973년 5월부터 9월까지 NewYork시에서 측정한 공기의 질과 관련하여 매일 측정한 자료로써 Ozone, Solar.R, Wind, Temp, Month, Day 총 6가지의 변수로 이루어져 있다. 여기서 날짜를 나타내는 Month와 Day를 제외한 나머지 4개의 변수를 분석에 이용하기로 한다. 자료의 출처는 New York State Department of Environmental Conservation에서 Ozone을, National Weather Service에서 나머지 기상 자료를 얻었다. Ozone은 13시 부터 15시 까지 Roosevelt Island에서의 오존의 농도이며 단위는 ppb이다. Solar.R은 Central Park의 주파수 대역 4000 ~ 7700 옹스트롬의 Langleys에서의 태양복사를 나타낸다. Wind는 LaGuardia 공항에서 07시 부터 10시까지의 평균 풍속을 나타내며 단위는 mph이다. Temp는 LaGuardia 공항에서 화씨의 일일 최고 온도를 나타낸다. Ozone이 반응변수이고 Solar.R, Wind, Temp는 설명변수이다. 총 관측된 자료의 수는 154개 이지만 결측치가 있는 개체들을 제외하였고 분석을 진행하던 중 극단적인 이상치 1개를 제외하고 총 110개의 개체를 가지고 분석을 진행하였다.

3. 자료 분석

각 변수들의 개별적인 특징을 알아보기 위해서 단변량 분석을 실시하였다. 그 결과가 아래의 Table 3.1에 제시되어 있다. 오존의 농도는 평균이 42.099이고 표준편차가 33.276, 태양 복사는 평균이 184.802이고 표준편차가 91.152, 평균 풍속은 평균이 9.94이고 표준편차가 3.558, 최고 온도는 평균이 77.793이고 표준편차가 9.53임을 알 수 있다.

Table 3.1 Univariate analysis

Variables	Mean	Variance	Standard deviation	Skewness	Kurtosis
Ozone	42.099	1107.290	33.276	1.265	1.316
Solar.R	184.802	8308.742	91.152	-0.492	-0.916
Wind	9.940	12.657	3.558	0.461	0.350
Temp	77.793	90.820	9.530	-0.228	-0.643

분석한 결과로는 수치적인 것은 쉽게 알 수 있으나 자료의 분포에 관한 특성은 뚜렷하지 않다. 단변량 분석 결과를 R 프로그램을 통해 히스토그램으로 시각화하여 나타낸 결과가 아래 Figure 3.1에 있다. 오존 농도를 보면 히스토그램의 형태가 오른쪽으로 꼬리가 긴 모양임을 알 수 있다. 이는 자료의 분포가 값이 작은 쪽으로 몰려 있다는 것을 알 수 있고 양의 왜도를 가지게 되는 것을 알 수 있다. 태양복사의 히스토그램의 형태는 왼쪽으로 꼬리가 긴 모양임을 알 수 있다. 하지만 히스토그램의 봉우리가 크지 않은 것으로 봐서 자료가 고르게 분포되면서 약간 큰 값이 많다는 것을 알 수 있다. 평균 풍속의 히스토그램은 종모양으로 평균에 대칭인 형태를 띠고 있고 최고온도의 히스토그램은 약간 오른쪽으로 치우친 모양임을 알 수 있다.

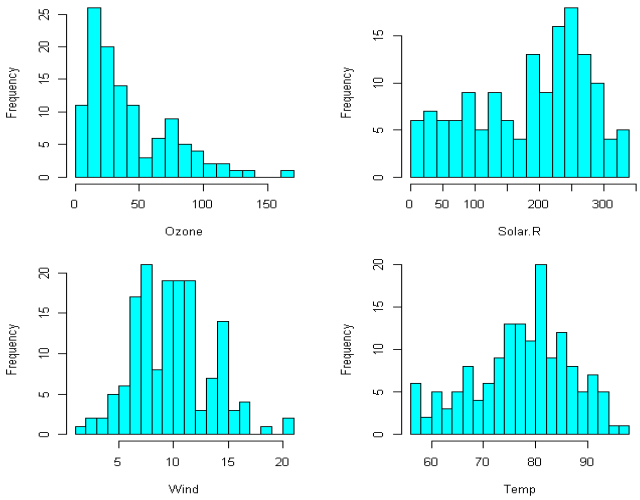

Figure 3.1 Univariate histogram

자료 “Airquality”는 태양 복사와 평균 풍속 및 최고 온도에 따른 오존의 농도를 나타낸 자료이다. 따라서 오존 농도에 대해서 다른 변수들이 각각 어떠한 관계를 가지고 있는지 단순 회귀 분석을 통해서 살펴보았다 (Table 3.2). 설명변수가 태양 복사 일 때 모형의 적합성도 유의하고 회귀계수도 유의하지만 결정계수가 0.1119으로 모형의 설명력이 상당히 떨어짐을 알 수 있다. 설명변수가 평균 풍속 일 때도 또한 모형의 적합성과 회귀계수 모두 유의하지만 결정계수가 0.3814로 낮고 설명력이 떨어짐을 알 수 있다. 설명변수가 최고 온도 일 때 마찬가지로 모형의 적합성과 회귀계수는 유의하지만 결정계수가 0.4809으로 모형의 설명력은 낮음을 알 수 있다. 태양 복사와 최고 온도의 회귀계수 추정치는 양수이므로 태양 복사 값과 최고 온도가 높을수록 오존의 농도는 높아지고 평균 풍속의 회귀계수 추정치는 음수이므로 풍속이 낮으면 오존의 농도가 낮아짐을 알 수 있다.

Table 3.2 Simple regression analysis

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr $>\|t\|$
Intercept	1	19.46958	6.91868	2.81	0.0058
Solar.R	1	0.12340	0.03345	3.69	0.0004
F Value: 13.61, p-value: 0.0004, R-Square: 0.1119					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr $>>\|t\|$
Intercept	1	99.50073	7.42195	13.41	<.0001
Wind	1	-5.73617	0.70289	-8.16	<.0001
F Value: 66.60, p-value: <.0001, R-Square: 0.3814					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr $>>\|t\|$
Intercept	1	-148.41637	19.22134	-7.72	<.0001
Temp	1	2.44844	0.24479	10.00	<.0001
F Value: 100.05, p-value: <.0001, R-Square: 0.4809					

단순 회귀 분석을 실시한 후 잔차에 대한 그림이 Figure 3.2에 있다. 위의 그림은 태양 복사에 대한 잔차 그림이다. 표준화 잔차의 값이 나팔모양으로 점점 커지는 것으로 보아 오차의 등분산성이 만족하지 않는다는 것을 알 수 있다. 왼쪽 아래의 그림은 평균 풍속에 대한 잔차 그림이다. 마찬가지로 태양 복사처럼 표준화 잔차의 값이 나팔모양으로 점점 커지고 있다. 이 또한 오차의 등분산성이 만족하지 않는 것을 알 수 있다. 오른쪽 아래의 그림은 최고 온도에 대한 잔차 그림이다. 마찬가지로 산점도의 모양이 일정하지 않아 적절한 변수변환이 필요하다는 것을 알 수 있다.

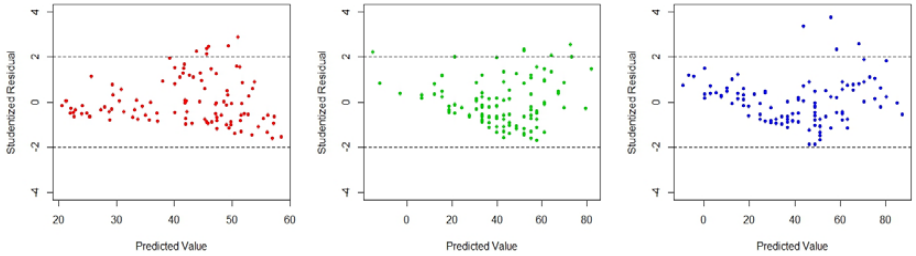

Figure 3.2 Residual diagram of simple regression analysis

아래의 Figure 3.3은 R을 이용하여 오존 농도와 다른 변수들 간의 산점도를 나타내었다. 먼저 오존 농도와 태양 복사의 산점도를 보면 태양 복사가 증가하면서 오존 농도의 분포가 점점 퍼지는 것을 알 수 있다. 따라서 위에서 확인하였던 결정계수가 상당히 작았던 이유를 산점도를 통해 알 수 있다. 평균 풍속과 최고 온도도 마찬가지로 자료가 선상에 있는 것이 아니라 점점 퍼지는 형태를 약하게 띠고 있다. 그래서 결정계수가 높지 않았던 것을 알 수 있다. 따라서 이 산점도와 잔차도를 통해 변수의 적절한 변환이 필요함을 알 수 있다.

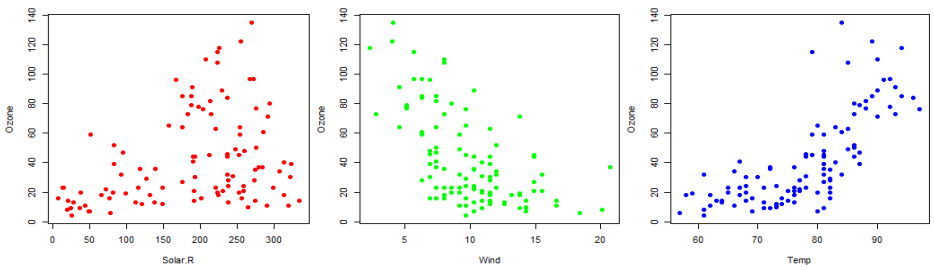

Figure 3.3 Scatter plot of simple regression analysis

따라서 Ozone을 로그변환을 하여 회귀분석을 실시해 보았다. Wind는 결정계수가 약간 감소했지만 Solar와 Temp는 단순회귀분석에 비해 확실히 결정계수가 증가함을 보여준다. 로그변환을 하는 것이 변환하기 전보다 모형의 설명력이 더욱 증가하였음을 알 수 있다. 표준화 회귀계수를 살펴보면 Temp의 회귀계수의 절대값이 가장 크고 Solar의 회귀 계수 절대값이 가장 낮다. 그리고 t Value를 보았을 때도 마찬가지 이다. 이는 오존 농도가 최고 온도에 가장 큰 영향을 받고 그 다음으로는 평균 풍속에 영향을 많이 받고 마지막으로 태양 복사가 가장 덜 영향을 미치는 것으로 보여진다.

Table 3.3 Simple regression analysis

Variable	DF	Parameter Estimate	StandardizedEstimate	Standard Error	t Value	Pr $>\|t\|$
Intercept	1	2.73921	0	0.16130	16.98	<.0001
Solar.R	1	0.00380	0.42420	0.00078	4.87	<.0001
F Value: 23.70, p-value: 0.0001, R-Square: 0.1799						
Variable	DF	Parameter Estimate	StandardizedEstimate	Standard Error	t Value	Pr$>\|t\|$
Intercept	1	4.79752	0	0.18268	26.26	<.0001
Wind	1	-0.13584	-0.60284	0.01730	-7.85	<.0001
F Value: 61.66, p-value: <.0001, R-Square: 0.3634						
Variable	DF	Parameter Estimate	StandardizedEstimate	Standard Error	t Value	Pr $>\|t\|$
Intercept	1	-1.49997	0	0.43485	-3.45	0.0008
Temp	1	0.06345	0.74071	0.00554	11.46	<.0001
F Value: 131.29, p-value: <.0001, R-Square: 0.5487						

Figure 3.4에 로그 변환 후 회귀분석에 대한 잔차 그림을 나타내어 보았다. 로그 변환하기 전의 Figure 3.2에 비해서 나팔모양도 사라지고 잔차가 고르게 분포되어 있는 것을 알 수 있다.

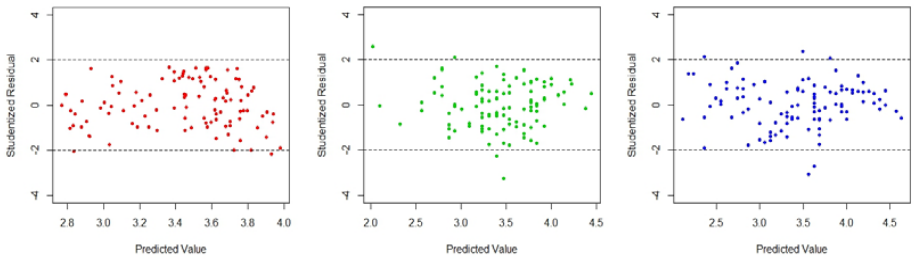

Figure 3.4 Residual diagram of simple regression analysis

아래 Figure 3.5는 산점도 행렬로써 네 변수들 간의 상관관계를 한눈에 파악 할 수 있는 시각화 방법이다. 산점도 행렬의 위쪽 패널에는 두 변수의 관계를 loess라고 불리는 비모수 회귀모형으로 추정하여 선으로 나타내었다. 아래쪽 패널은 산점도에 선형 회귀를 나타내었다. 대각선 패널은 히스토그램을 나타냈다. 산점도행렬에는 다양한 방법으로 자료를 나타낼 수 있다. 패널만 자기가 필요한 부분을 설정만 하면 원하는 형태를 입력할 수 있다. 아래의 그림으로 오존과 로그변환 한 오존을 비교할 수 있다.

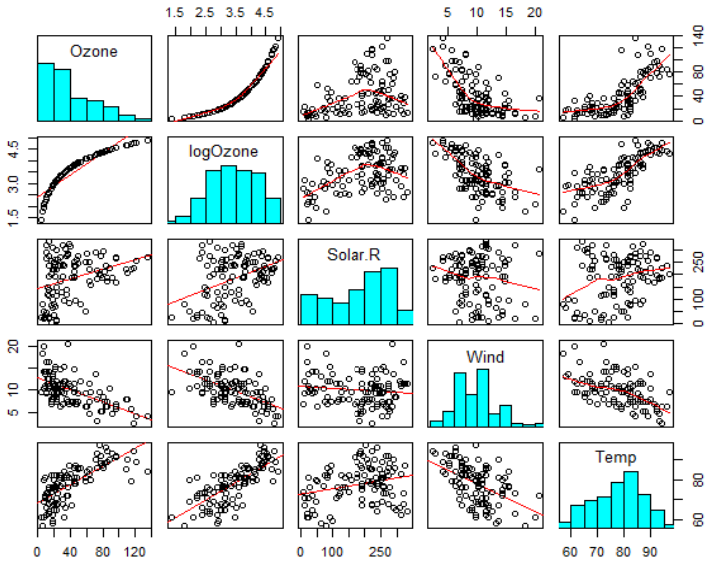

Figure 3.5 Scatter plot matrix

단순 회귀 분석을 통해 각 변수가 오존 농도와 어떤 관계가 있는지 파악을 해 보았고 이번에는 다중 회귀 분석을 실시하여 여러 변수와 오존 농도와의 관계를 살펴보았다. 그 결과가 아래 Table 3.4이다. 모형의 적합성 검정은 통계적으로 유의하였다. 또한 회귀계수도 모두 통계적으로 유의하고 결정계수는 0.6736이다. 단순 회귀 분석을 실시했을 때 보다 모형의 설명력이 훨씬 높은 것을 알 수 있다. 그리고 다중공선성을 체크하기 위하여 VIF를 살펴보면 모두 10보다 작으므로 다중공선성은 존재하지 않는다고 할 수 있다. 세 변수 중 가장 영향력이 높은 순서를 살펴보면 절대값이 가장 큰 최고 온도가 가장 영향력이 크고 그 다음으로 영향을 미치는 것은 평균 풍속임을 알 수 있다. 태양 복사는 가장 영향력이 적은 것을 알 수 있다. 온도가 높을수록, 바람이 적을수록, 태양 복사량이 높을수록 오존의 농도는 커지는 것을 알 수 있다. 그리고 아래에는 잔차그림을 나타내었다. 잔차도의 모양이 고른 것으로 보아 오차항에도 문제가 없는 것으로 판단할 수 있다.

Table 3.4 Multiple regression analysis

Variable	DF	Parameter Estimate	Standardized Estimate	Standard Error	t Value	Pr> $\|t\|$	VIF
Intercept	1	0.26144	0	0.52050	0.50	0.6165	0
Solar.R	1	0.00219	0.24472	0.00052	4.25	<.0001	1.07769
Wind	1	-0.06928	-0.30746	0.01451	-4.77	<.0001	1.34706
Temp	1	0.04446	0.51897	0.00568	7.83	<.0001	1.42687
F Value: 72.91, p-value: <.0001, R-Square: 0.6736							

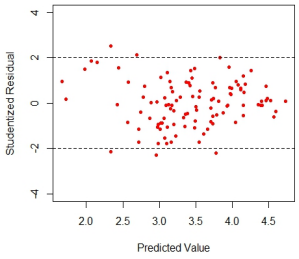

Figure 3.6 Residual diagram of multiple regression analysis

4. 데이터 시각화

자료가 어떤 특성을 가지고 있는지 쉽게 알아보기 위해서 여러 가지 방법으로 시각화를 해보았다. 변환점을 찾는 많은 통계적 연구가 있지만 (Zeileis 등, 2002), 시각적으로 자료를 표현하는 것은 시각적 통찰력이라는 특수한 능력을 활용하여, 숨겨진 채 계속 못 찾을 수 있는 의미있는 패턴을 발견하게 한다 (Stephen, 2007). 이에 근거하여, 먼저 자료를 Figure 3.5의 산점도행렬에서 loess로 추정된 선에서 변곡하는 지점을 시각적으로 판단하여 각 설명변수를 이분화 하였다. 이분화 하는 기준은 태양 복사는 200, 평균 풍속은 10, 최고 온도는 78을 기준으로 이분화를 하였다. 이분화된 데이터를 가지고 로그변환을 한 오존 농도에 대해 상자그림을 그린 것이 아래의 Figure 4.1이다. T는 Temp인 최고 온도를, W는 Wind인 평균 풍속을, S는 Solar인 태양 복사를 나타낸다. 상자그림을 보면 단순회귀분석에서 살펴보았듯이 태양 복사는 높을수록, 평균 풍속은 낮을수록, 최고 온도는 높을수록 오존의 농도가 높은 편임을 알 수 있다. 세 가지 요인 중에서 다중회귀분석에서 살펴보았던 대로 온도가 오존의 농도에 가장 큰 영향을 미치고 그 다음으로는 풍속, 마지막으로 태양 복사가 가장 덜 영향을 미치는 것을 알 수 있다

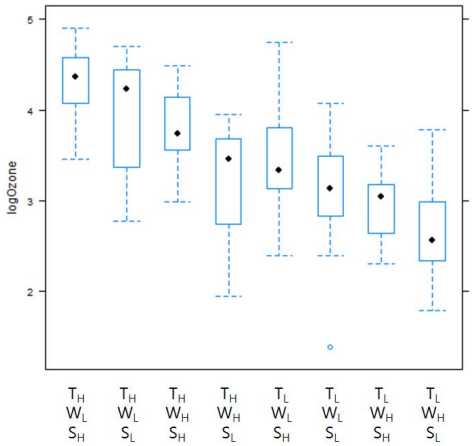

Figure 4.1 The box plot of bipartite independent variables

어떤 요인이 더 오존 농도에 영향을 많이 미치는지 더 알기 쉽게 보기 위해서 3차원으로 산점도를 그려보았다. 3차원 산점도는 드래그를 함으로써 각도를 자유자재로 변경할 수 있다. 점을 이분화하여 나타내었고 색상이 선명하면 더 가까이에 있는 것이고 흐릿하면 더 멀리 있는 것이다. 그리고 점들의 분포를 대략적으로 구의 형태로 나타내어 보았다. 위의 두 개의 산점도는 최고 온도를 이분화 하였고, 가운데의 두 개의 산점도는 평균 풍속을 이분화 한 것이다. 마지막 두 개의 산점도는 태양 복사를 이분화 한 것이다. 위의 산점도의 파란색은 온도가 낮은 부분이고 빨간색은 온도가 높은 부분이다. 살펴보면 그룹이 확실히 나눠짐을 알 수 있고, 풍속이 높으면서 온도가 낮을수록 오존의 농도는 낮고 풍속이 낮으면서 온도가 높을수록 오존의 농도는 높다는 것을 알 수 있다. 그리고 위의 상자그림에서도 보았듯이 오른쪽의 산점도로 보아 태양 복사는 온도와는 크게 관련이 없는 형태임을 알 수 있다. 가운데 두 개의 산점도에서는 초록색이 풍속이 낮은 부분이고 주황색이 풍속이 높은 부분을 나타낸다. 온도보다는 확연하지 않지만 대략적으로 풍속이 낮은 부분이 오존 농도가 높고 풍속이 높은 부분이 오존 농도가 낮음을 보여준다. 위의 두 개의 산점도와 마찬가지로 온도가 높으면서 풍속이 낮으면 오존의 농도가 더욱 높아지는 형태를 보여준다. 또한 태양 복사는 풍속과는 크게 관련이 없는 형태를 보여준다. 마지막 산점도는 태양 복사가 다른 요인들과는 영향을 별로 미치지 않은 것을 자명하게 나타내고 있다. 즉, 다중회귀분석에서 살펴보았던 결과와 동일한 영향력을 확인할 수 있다.

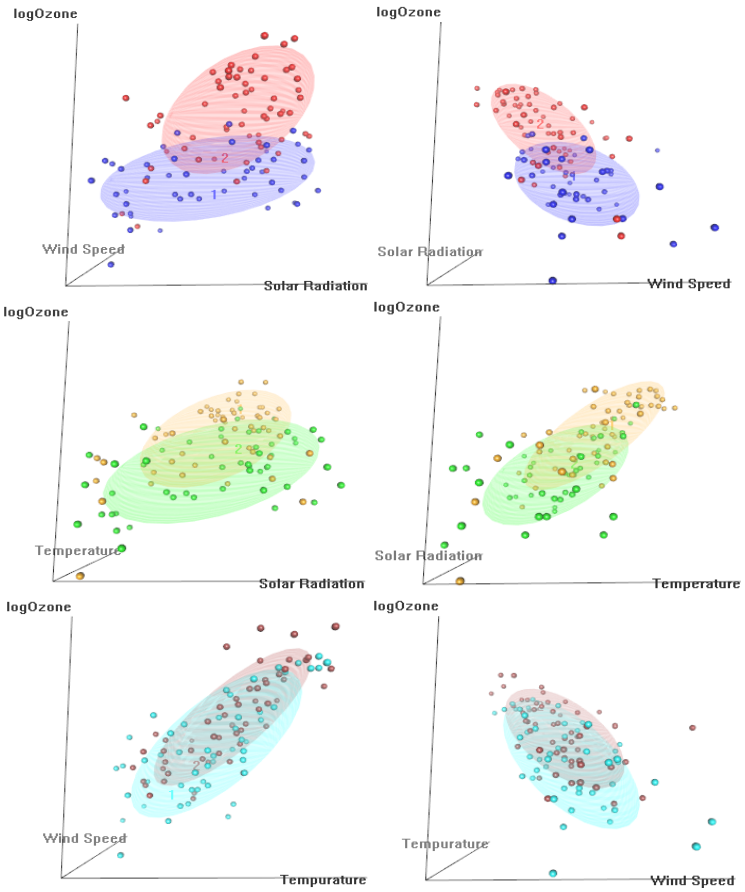

Figure 4.2 Three-dimensional scatter plot

아래의 그림은 Table 3.4의 회귀모형을 3차원 투시도에 적합 시켜 나타낸 것이다. 왼쪽과 오른쪽으로 나눈 것은 자료를 이분화 하였을 때 작은 부분을 왼쪽으로 큰 부분을 오른쪽에 나타내었다. 먼저 태양 복사에 대한 투시도는 좌우의 모양이 큰 차이가 없는 것으로 봐서 태양 복사의 변화에 대해서 온도와 풍속이 함께 오존의 농도에 크게 영향을 주지 않는 것을 보여준다. 주로 온도에 의해 오존의 농도가 크게 좌우됨을 알 수 있다. 온도를 나타낸 투시도를 보면 더욱 확실히 알 수 있다. 온도가 낮은 부분은 어떠한 요인으로도 오존의 농도는 크게 변함이 없는 것이 보이지만, 온도가 높은 부분에서는 오존의 농도가 전체적으로 크게 증가를 하였고, 풍속에 의해서만 약간 영향을 받는 것을 보였다. 풍속을 나타내는 투시도를 보면 왼쪽과 오른쪽에 대해서 온도의 변화에 대해 오존 농도의 차이가 있었는데 풍속이 높은 경우는 태양 복사가 낮으면 온도가 아무리 높다고 하더라도 오존의 농도는 상당히 낮음을 알 수 있다. 이렇게 투시도를 통해서 더 자세하게 요인들이 오존의 농도에 어떠한 효과를 주는지를 확인할 수 있었다.

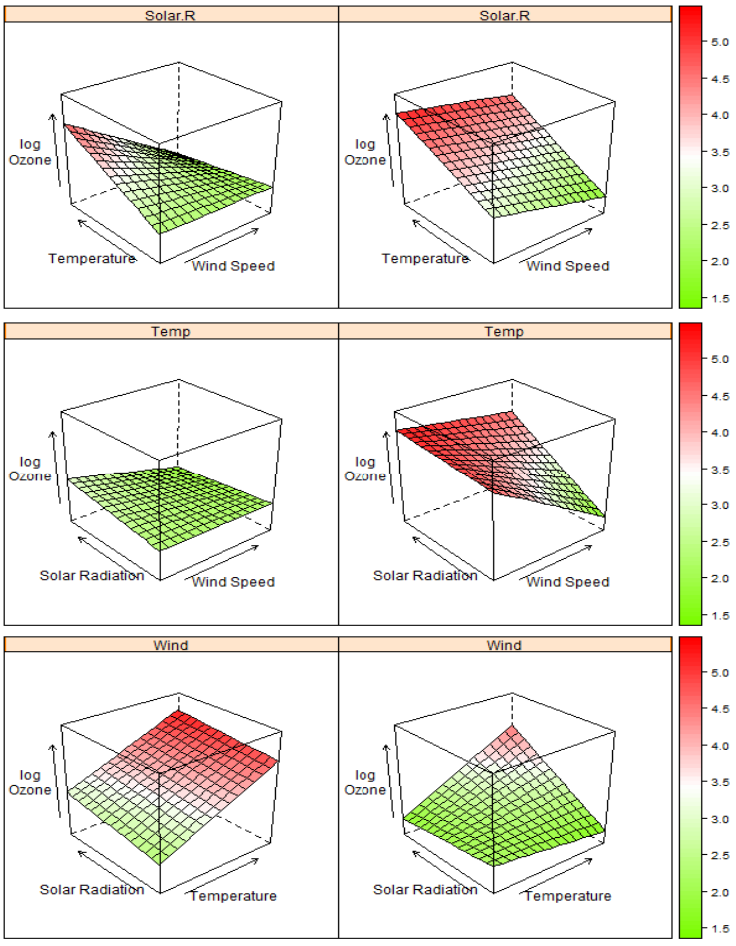

Figure 4.3 hree-dimensional perspective drawing

5. 결론

본 논문은 데이터 시각화를 통해 자료의 패턴을 파악하여 한눈에 자료의 특징을 알 수 있게 하고자 R을 이용하여 여러 가지 기법으로 그래프를 구현해 보았다. 2차원의 히스토그램과 산점도, 상자그림, 3차원 산점도와 투시도 등 다양한 방법으로 그래프를 구현하였고 이를 통해서 오존농도와 설명변수들 간에 어떠한 관련성이 있는지를 더욱 쉽게 파악해볼 수 있었다. 단순히 회귀분석 결과로만 보았을 때 알 수 없었던 특징들을 R을 이용한 데이터 시각화를 통해 보이지 않았던 자료의 특징들에 대해서도 파악할 수 있었고 자료를 있는 그대로만 해석하는 것이 아니라 변수들의 변환을 통해 다양한 방법으로 그래프를 그려보고 데이터를 시각화 하는 방법이 수많은 방법이 있다는 것을 알게 되었다. 데이터 시각화를 너무 어렵게 생각하기보다 R을 이용하면 누구나 쉽게 구현할 수 있고, 무료 패키지 프로그램인 R은 지금도 개발 중에 있으므로 앞으로는 더욱 다양한 데이터 시각화가 구현되라고 본다.

References

Bae, D., Park, H. and Oh, K. (2013). Big data trends and policy implications. *International Telecommunications Policy Review*, **25**, 37-74.

Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983). *Graphical Methods for Data Analysis*, Wadsworth.

Cho, J. (2014). Analysis of employee's characteristic using data visualization. *Journal of the Korean Data & Information Science Society*, **25**, 727-736.

Choi, K., Ham, Y. and Kim, S. (2013). Big data visualization. *KSCI Review*, **21**, 33-43.

Joo, S., Jung, J. and Ryu, K. (2013). Big data technology trends, visualizations of big and public data. *Smart Media Journal*, **2**, 37-43.

Kim, K. and Lee, K. (2014). A web application for open data visualization using R. *Journal of the Korean Association of Geographic Information Studies*, **17**, 72-81.

Lim, Y., Baek, S. and Yeon, S. (2012). Choice and focus for competitiveness of big data era. *Nurimedia Korean Studies Journals*, **29**, 3-10.

Park, D. (2007). Teaching statistical graphics using R. *The Korean Journal of Applied Statistics*, **20**, 619-634.

Park, S. (2014). Visualization and interpretation of cancer data using linked micromap plots. *Journal of the Korean Data & Information Science Society*, **25**, 1531-1538.

Stephen, F. (2007). Visualizing change : an innovation in time-series analysis. *Visual Business Intelligence Newsletter*, September 2007.

Zeileis, A., Leisch, F., Hornik, K. and Kleiber, C. (2002). Strucchange: an R package for testing for structural change in linear regression models. *Journal of Statistical Software*, **7**, 1-38.

Journal of the Korean Data & Information Science Society
2015, **26**(2), 399–408

http://dx.doi.org/10.7465/jkdi.2015.26.2.399
한국데이터정보과학회지

Data visualization of airquality data using R software[†]

Youngchang Oh[1] · Eunsik Park[2]

[12]Department of Statistics, Chonnam University

Received 3 February 2015, revised 12 February 2015, accepted 19 March 2015

Abstract

This paper presented airquality data through data visualization in several ways and described its characteristics related to statistical methods for analysis. Software R was used for visualization tools. The airquality data was measured in New York city from May to September of year 1973. First, simple, exploratory data analysis was done in terms of both data visualization and analysis to find out univariate characteristics. Then through data transformation and multiple regression analysis, model for describing the airquality level was found. Also, after some data categorization, overall feature of the data was explored using box plot and three-dimensional perspective drawing and scatter plot.

Keywords: Box plot, data visualization, scatter plot matrix, three-dimensional perspective drawing, three-dimensional scatter plot.

[†] This study was financially supported by Chonnam National University, 2014.
[1] Graduate school master course, Department of Statistics, Chonnam National University, Gwangju 500-757, Korea.
[2] Corresponding author: Professor, Department of Statistics, Chonnam National University, Gwangju 500-757, Korea. E-mail: espark02@jnu.ac.kr