사고다발지역의 특정현황 포착을 위한 경기도 사고 데이터 군집분석

손정습*, 이동건**

Jeongseup Son*, Dong-geon Lee**

ABSTRACT

본 연구는 사고다발지역의 유사한 특성을 파악하여 사고다발지역의 사망사건의 주된 요인을 파악하는데 목적이 있다. 해당 목적을 통해, 기존 교통사고의 개선안의 문제점을 지적한다. 이를 통해, 교통사고 발생 시 주된 사망요인과 긴밀한 연관이 있는 군집을 바탕으로 교통사고 개선을 할 것을 제시한다. 또한, 주요 군집의 교통사고 개선은 사망확률이 높은 교통사고를 포착할 뿐만 아니라, 나아가 연관된 교통사고 발생 시 즉각적으로 구급활동이 이루어질 수 있도록 조취를 취하게 한다.

Key Words : 교통사고, 사고다발지역, 군집분석

* 인천대학교 산업경영공학과 학부 3학년
** 인천대학교 산업경영공학과 학부 3학년

1. 서론

우리나라는 2016년 OECD회원국 도로 사망자가 많은 나라 중 3위를 기록하고 있다. 해당 순위는 대략 10만 명 당 사망자 수가 8.38명 되는 것을 의미한다.[1] 우리나라의 교통사고는 2008년도에 실시한 '교통사고 사상자 절반 줄이기' 운동 이후 상당히 크게 줄었으나, 최근 다시 교통사고 사망자 수가 크게 증가하고 있으면 이 문제가 대두된다. 정부의 지속적인 노력으로 사망자가 감소고 있는 추세이긴 하나 일부분의 문제를 개선한 것으로 보인다. 차트를 보면 알 수가 있다.

◐ 사망자 수 3,781명, 9.7% 감소(지난 10년간 총 35.6% 감소)

구분	2017년	2018년	증감률
사고 건수	21만6,335건	21만7,148건	0.4%⇧
사망자 수	4,185명	3,781명	9.7%⇩
부상자 수	32만2,829명	32만3,036명	0.1%⇧

※ 음주운전(21.2%), 어린이(37.0%), 보행자(11.2%) 사망자 크게 감소

< 차트1. 연간 사망자 수 감소 >

가장 크게 감소한 부분은 어린이부분이지만, 실제 사고다발 시 어린이 사고의 사망률이 높은지는 의문이다. 그리고 보행자의 사망률은 11.2%정도로 감소하였으나, 실제 사고 상태별 비중에서는 보행자의 경우 가장 높은 상태이기 때문에 앞으로도 지속적인 개선이 필요한 상황이다. 쉽게 일반적인 경우를 생각해 볼 수 있다. 경험에 입어 생각을 해보면 현 우리 사회에 주위에서 쉽게 교통준규 습관에 대해 유추가 가능하다. 등하교 길에도 굉장히 높은 비율로 여러 보행자들의 불안전한 교통규칙 모습이 들어난다. 이런 문제에 대해서는 사회 일반적으로 지적이 가능할 정도의 문제로 인식이 가능하다. 특히, 신호위반에 대해 일반적 대중들은 굉장히 둔감해 보이는 경향이 있다. 이를 통해, 현 교통사고 사망률의 지속적인 유지와 개선을 위해 주의해야 하는 사고 유형에 대해 인식할 필요가 있으며, 사망률이 높은 사고에 대해 즉각 응급처치가 동원될 수 있도록 해야 한다.

따라서, 본 프로젝트는 사고다발지역의 통계 데이터를 사망자 발생을 기준으로 분석하여 특정 군집으로 분류시키는 군집분석을 통해 사고다발지역상의 사망발생조건 포착하고 이를 통해 실질적으로 개선시켜야 할 문제를 지적과 개선방향을 제시하고자 한다.

2. 데이터 배경

2.1 데이터 관찰

데이터는 경기도 데이터라는 공공 홈페이지를 통해 확보하였다. 경기도 내의 시, 군 별 사고에 대하여 통계적으로 정리되어있는 csv파일이다. 데이터는 총 17개의 요인으로 구성되어져있다.(시, 군명, 사고년도, 사고유형구분, 다발지식별자, 다발지역코드, 법정 동코드, 위치코드, 시도시군구명, 사고지역위치명, 발생건수, 사상자수, 사망자수, 중상자수, 부상자수, 위도, 경도) 해당 데이터 요인은 공공 홈페이지 상에서 쉽게 얻을 수 있는 데이터이다. 지역적 정보와 사고유형 그리고 사고의 발생요인들로만 이를 군집화하고자 한다.

2.2 데이터 요인 추가

1) 인근 신고센터 거리

기존 데이터는 시, 군명뿐만 아니라 구체적 해당 사고의 위도와 경도가 포함되어있다. 이를 활용하여 사고 발생 시 가장 가깝게 신고 및 접수 그리고 대처까지 가능한 센터들을 기준으로 해당 사고 발생지점과 센터와의 거리를 새로운 요인으로 추가하였다. 데이터 병합 시에는 소방, 경찰, 지구대, 치안센터 등의 현황을 시, 군명을 key로 하여 기존 데이터에 병합을 하였다.

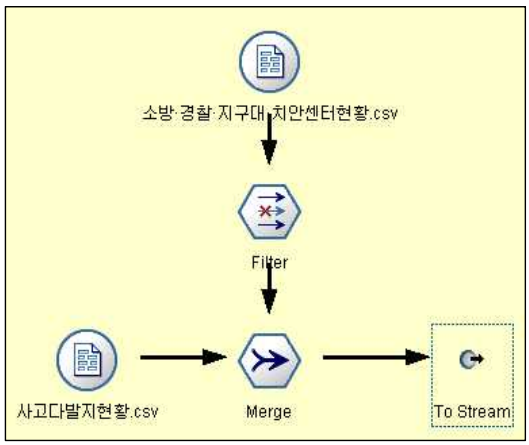


그림 1. 데이터 병합 과정

데이터 병합 후, 엑셀 파일 옮긴다. 그리고 사고 발생지와 센터간의 거리를 엑셀 함수를 통해 거리계산식[1)]을 바탕으로 계산한다.

1) 거리계산식 :
=ACOS(COS(RADIANS(90-지점1의위도))*COS(RADIANS(90-지점2의위도))+SIN(RADIANS(90-지점1의위도))*SIN(RADIANS(90-지점2의위도))*COS(RADIANS(지점1의경도-지점2의경도)))*6371

2) 교통 안전 지수

지역별로 나뉘어져 있는 데이터에 교통 안전 지수를 추가적인 요인으로 포함시킨다. 교통 안전 지수는 전국 기초자치단체를 대상으로 교통사고 심각도별 사고건수와 사상자수를 기초로 인구와 도로연장을 고려하여 지자체별 교통안전 수준을 평가한 지수이다.

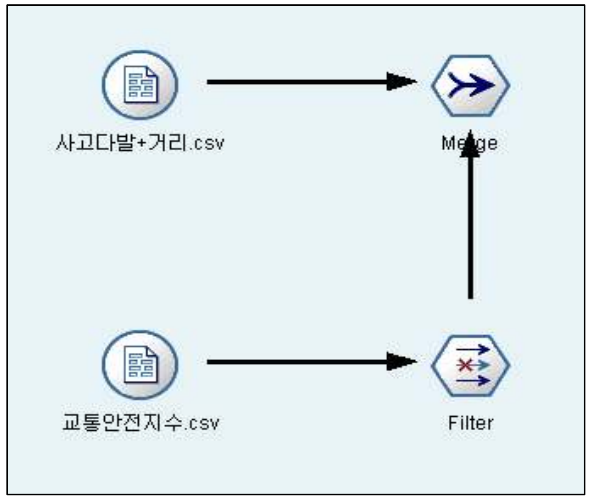


그림 2. 데이터 병합 과정

위의 과정을 거쳐 지역별 교통안전지수를 살펴볼 수 있었다.

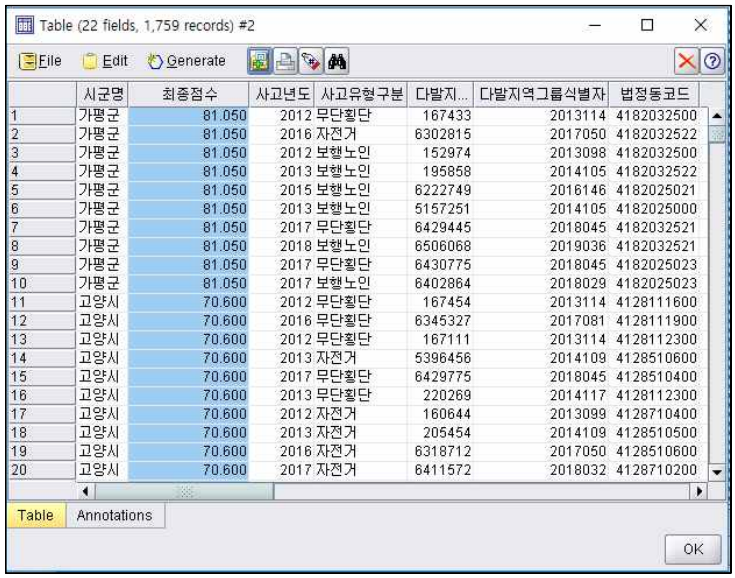


Table (22 fields, 1,759 records) #2

File Edit Generate

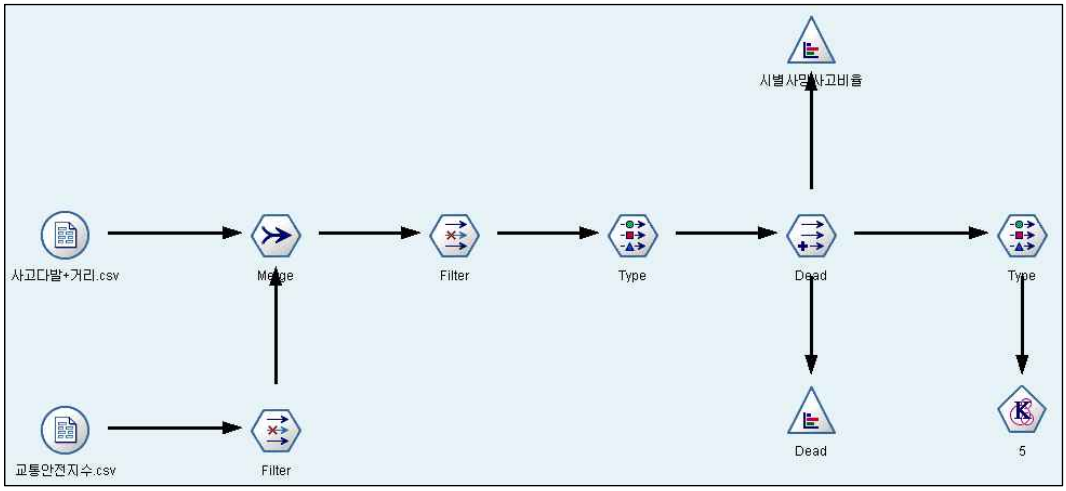
	시군명	최종점수	사고년도	사고유형구분	다발지...	다발지역그룹식별자	법정동코드
1	가평군	81.050	2012	무단횡단	167433	2013114	4182032500
2	가평군	81.050	2016	자전거	6302815	2017050	4182032522
3	가평군	81.050	2012	보행노인	152974	2013098	4182032500
4	가평군	81.050	2013	보행노인	195858	2014105	4182032522
5	가평군	81.050	2015	보행노인	6222749	2016146	4182025021
6	가평군	81.050	2013	보행노인	5157251	2014105	4182025000
7	가평군	81.050	2017	무단횡단	6429445	2018045	4182032521
8	가평군	81.050	2018	보행노인	6506068	2019036	4182032521
9	가평군	81.050	2017	무단횡단	6430775	2018045	4182025023
10	가평군	81.050	2017	보행노인	6402864	2018029	4182025023
11	고양시	70.600	2012	무단횡단	167454	2013114	4128111600
12	고양시	70.600	2016	무단횡단	6345327	2017081	4128111900
13	고양시	70.600	2012	무단횡단	167111	2013114	4128112300
14	고양시	70.600	2013	자전거	5396456	2014109	4128510600
15	고양시	70.600	2017	무단횡단	6429775	2018045	4128510400
16	고양시	70.600	2013	무단횡단	220269	2014117	4128112300
17	고양시	70.600	2012	자전거	160644	2013099	4128710400
18	고양시	70.600	2013	자전거	205454	2014109	4128510500
19	고양시	70.600	2016	자전거	6318712	2017050	4128510600
20	고양시	70.600	2017	자전거	6411572	2018032	4128710200

Table Annotations

OK

그림 3. 지역별 교통안전 지수

3) 불필요 요인의 제거와 타겟 설정



Filter노드를 통해 다발지식별자, 자발지역그룹식별자, 법정동코드, 위치코드, 시도시군구명, 사고지역위치명, caseid, 인근긴급처위도, 인근긴급처경도와 같은 불필요한 데이터와 중복되는 요인들을 제거하였다.

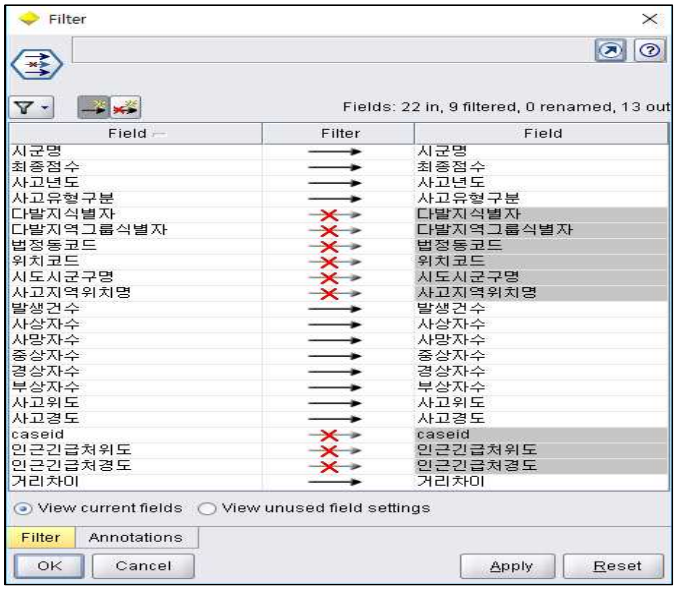


그림 4. 불필요한 요인 제거

Type노드를 이용하여 시군명, 최종점수, 사고년도, 사고유형구분, 발생건수, 사상자수, 사망자수, 중상자수, 경상자수, 부상자수, 사고위도, 사고경도, 거리차이의 Type을 설정하고 타겟을

설정하였다.

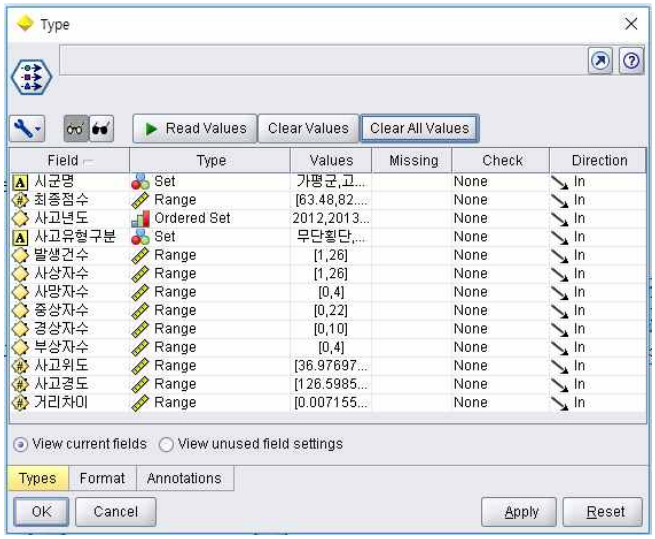


그림 5. 타겟설정

또한, 우리는 발생했던 사고가 몇 명이 사망한 사고인가를 궁금한 것이 아니고 발생한 사고가 사망사고인지 아닌지의 여부로 데이터모델링을 진행할 것이므로 사망자수의 Values를 [0,4]에서 사망자가 발생 했는지 True와 False 값으로 나타내도록 설정하였다.

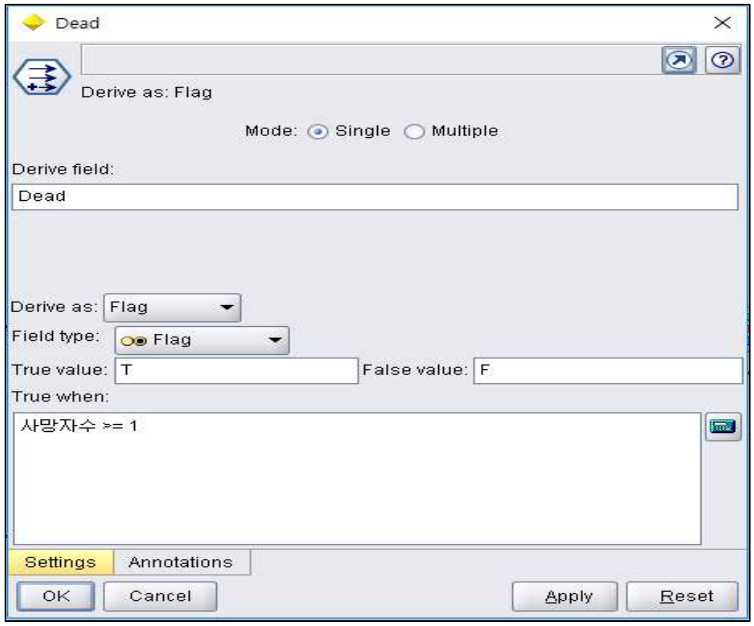


그림 6. 사망자수의 Values 변경

2.3 데이터 시각화

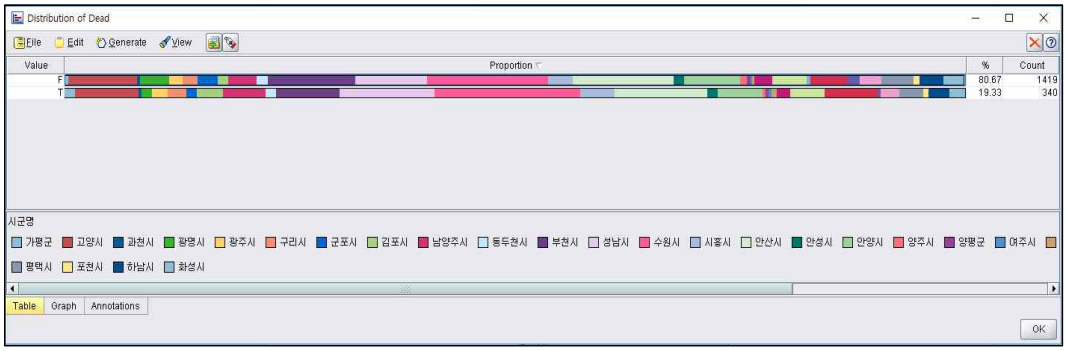


그림 7. 교통사고 비율

그림7을 통해 총 1759건의 교통사고에서 사망사고는 340건으로 19.33%나 차지하고 있는 것을 살펴 볼 수 있었고 수원시가 시별 교통사고 비율중 제일 큰 비중은 차지하고 있음을 알 수 있다.

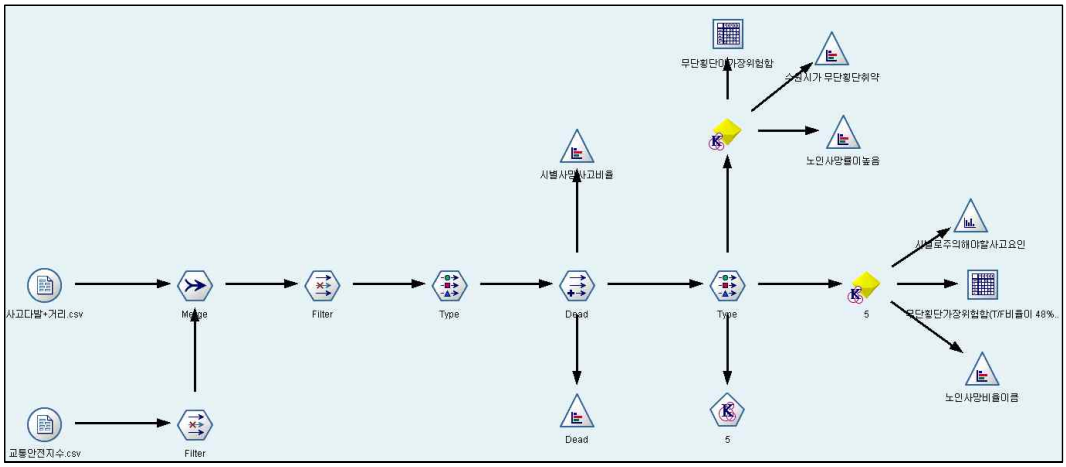
그림 8. 시별 사망사고 비율

하지만, 그림7과 그림8을 비교해서 살펴보면 각 시별 교통사고의 비율과 해당 시의 사망사고의 비율이 다르다는 것을 알 수 있다. 수원시에서 교통사고가 가장 많이 일어나지만 수원시의 교통사고 중 사망사고의 비율이 상위권에 위치하고 있지 않다. 즉, 교통사고의 발생건수가 많다고 하여 해당 시의 사망사고의 비율 또한 높다고 할 수 없는 것이다.

3. 군집분석

3.1 군집화 과정

본 연구는 위의 과정처럼 경기도의 사고 데이터들을 분석하였고 여러 시사점을 얻을 수 있다. 해당 데이터들을 모델링 과정을 통해 어떠한 요인들이 사망사고로 이어지게 되는지 살피고 이를 통해, 각 시에서 주의해야할 요인들을 찾아낼 것 이다. 방법은 K-Means 모델링을 사용했으며 5개의 cluster를 만들어 내도록 하였다. 5개의 모델로 설정하였을 때 가장 적합한 결과를 얻을 수 있었기 때문이다.



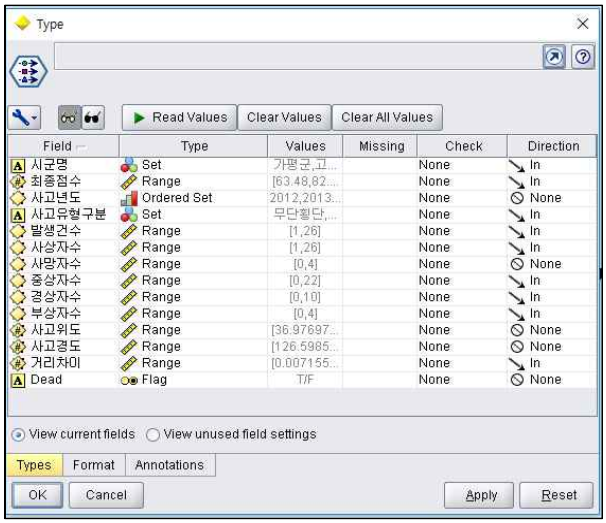


그림 9. Type노드

그림과 같이 모델링을 진행하기 전 다시 Type노드를 이용하여 사고년도, 사망자수, 사고위도, 사고경도, Dead를 제외 하였다. 사고위도와 사고경도는 시군명이라는 위치적 요인으로 대체하여 분석될 수 있으며, Dead는 최종 분석결과와 비교하기 위해 제외하였다.

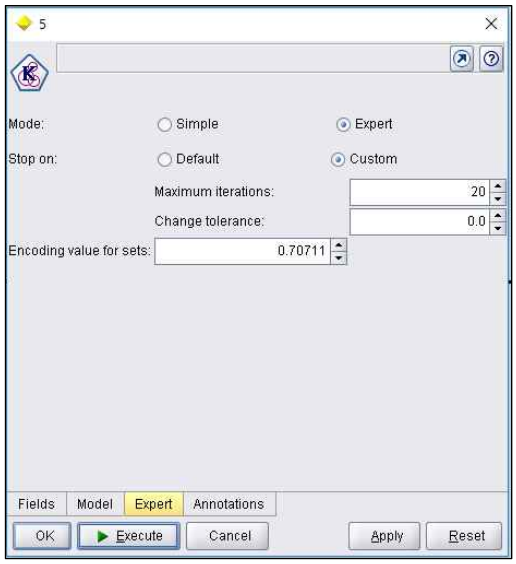


그림 10. Model Expert Setting

본 연구는 방법론적으로, K-Means 모델을 사용하였고 Expert에서 Mode는 Expert, Stop on은 직접 설정하였으며 Maximum iterations는 20, Change tolerance는 0으로 설정하였다. Encoding value는 0.770711로 설정하여 모델링을 진행하였다.

4. 결과 분석

4.1 모델링

그림 11. 군집 별 사망 분포

위 그림과 같이 사망사고를 5개의 cluster로 분할하였을 때, cluster-4의 비중이 가장 높은 것을 볼 수 있다. 따라서 cluster-4의 속성을 살펴보면 어떤 요인들이 사망사고에 가장 큰 영향을 주는지를 알 수 있을 것이라고 생각했다. 하지만, 위의 데이터 시각화 과정에서 수원시를 예를 들어보면 전체 교통사고 건수가 많다고 하여 해당 시의 사망사고의 비율이 큰 것은 아닌 것을 볼 수 있다. 다음으로 Dead의 Matrix를 살펴보았다.

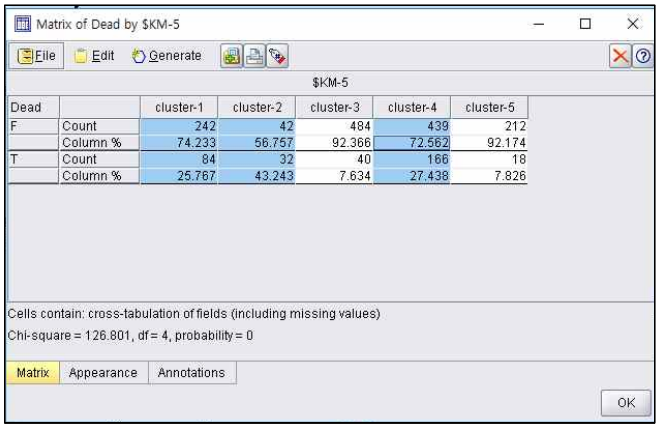
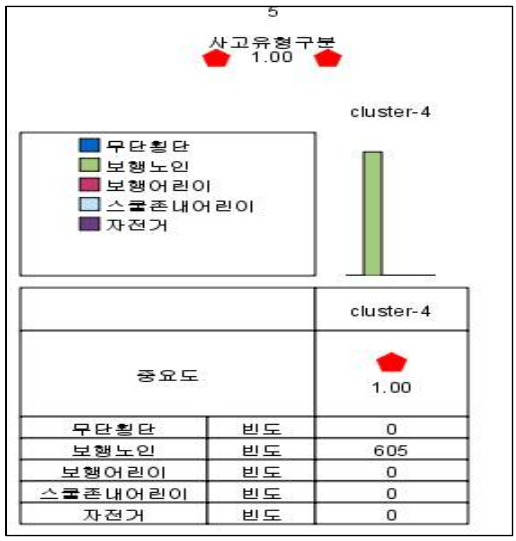


그림 12. 군집과 사망사건의 매트릭스 표

위 표를 살펴보면 cluster-2의 사고발생건수와 사망사고 건수는 낮은 수치를 보여주고 있지만, 해당 cluster 내에서의 교통사고가 발생하였을 때 사망사고의 비중이 약 43%를 차지하고 있다. 이것은 해당 cluster로 분할된 사고가 발생하였을 때 43% 확률로 사망한다는 것 이다. 따라서, 사고가 발생하였을 때 사망사고로 연결되는 비중이 큰 cluster 또한 고려해야 할 사항으로 생각할 수 있다.

4.2 Cluster 살펴보기



		cluster-4
중요도		1.00
무단횡단	빈도	0
보행노인	빈도	605
보행어린이	빈도	0
스쿨존내어린이	빈도	0
자전거	빈도	0

그림 13

위 그림을 통해 각 cluster의 속성을 알아낼 수 있었으며 cluster-4는 보행노인으로 볼 수 있고 교통사고와 사망사고에 큰 영향을 끼치는 것을 알 수 있다.

4.3 모델링의 적용

본 연구는 모델링을 통해 찾아낸 요인들을 이용하여 각 지역에 적용하였을 때, 예를 들어 각 지역에서 어느 연령층에 초점을 맞춰서 사고 대비를 위한 안전교육 해야 하는지, 어떠한 발생 원인을 주의 깊게 통제해야 하는지 등을 미리 준비할 수 있게 5개의 지역을 살펴보며 그 방법을 제시해보았다.

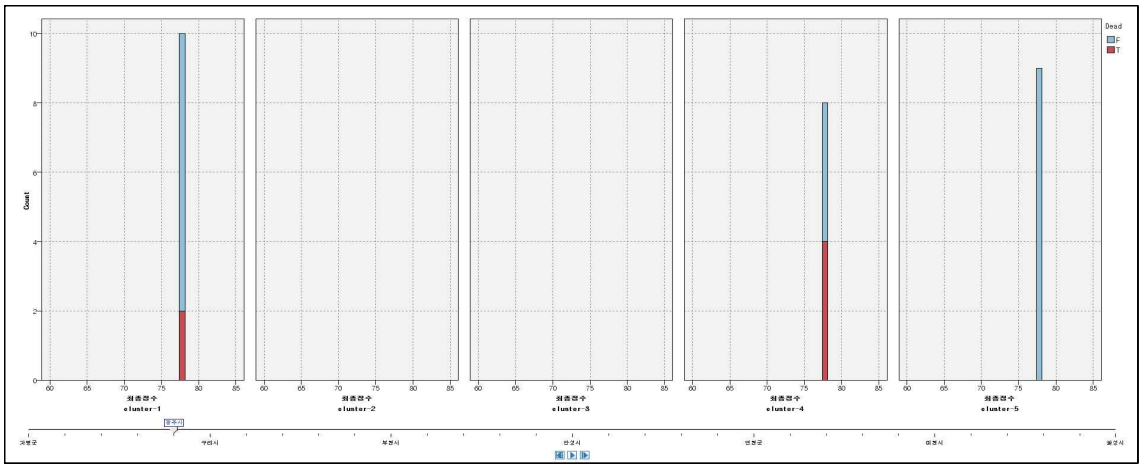


그림 14. 광주시 교통안전지수

위 그래프를 통해 광주시는 사망사고가 cluster-1과 cluster-4에서 나타나고 있다. 즉 스쿨존 내 어린이들과 노인들의 사고가 사망사고로 연결되는 경우가 많이 발생하는 것이므로 예를 들어 노인들과 을 겨냥한 안전교육, 스쿨존 내 안전운전에 대한 교육 등 광주시에서는 어디에 초점을 맞춰 사고를 대비할 수 있는지 알 수 있다.

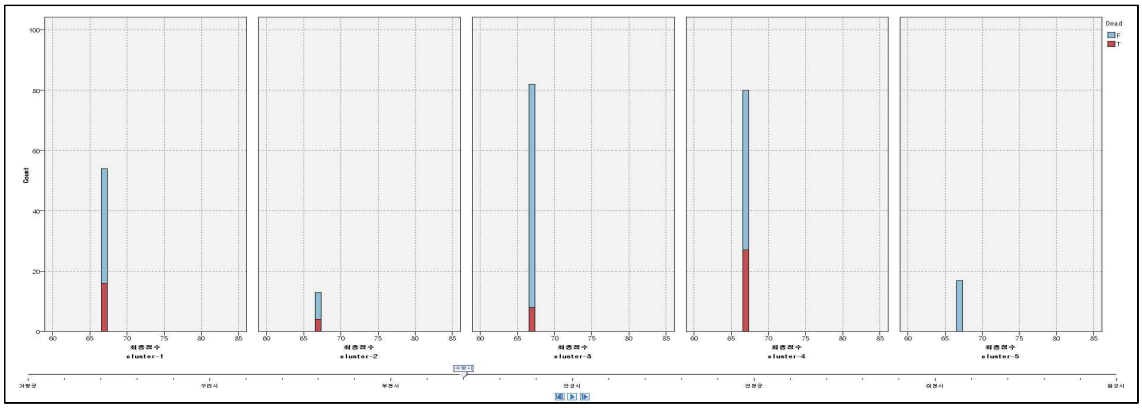
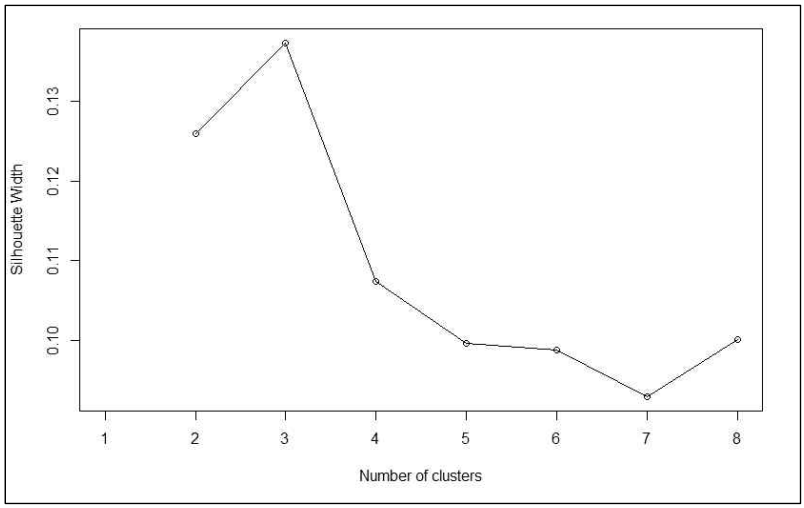


그림 15. 수원시 교통안전지수

위 그래프를 살펴보면 수원시에서는 사망사고가 cluster1부터 cluster4까지에 이르러 나타나는 것을 볼 수 있다. 따라서 각 사망사고가 발생한 cluster들을 고려하여 사망사고를 대비할 수 있을 것이고 어떠한 요인에 비중을 둬서 사고대비를 실시해야 하는지도 정할 수 있을 것 이다.

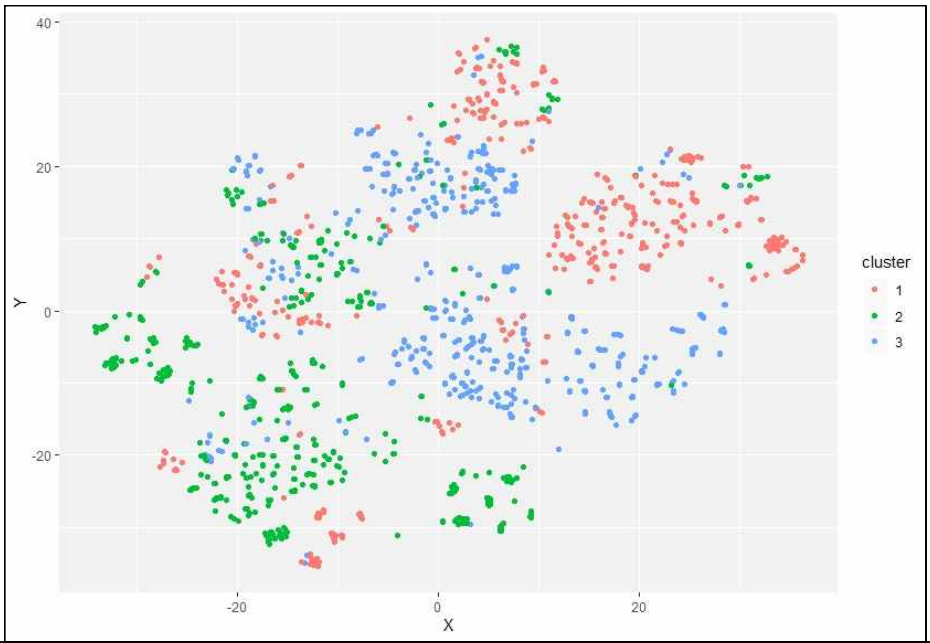
4.4 R 결과

해당 K-Means 모델을 R로 실행시켜보았다. 기존의 K-Means 모델은 카테고리형 데이터는 수용하기 힘들다. 따라서, Mixed type data clustering function을 사용하였다. 해당 방법은 R cluster package상의 Gower distance를 사용하는 것 이다. Gower distance를 통하여 범주형 데이터 또한 자동적으로 해당 특징을 0부터 1사이의 정규화를 시켜 거리간의 차이를 통한 K-Means 모델을 가능하게 한다. R결과는 다음과 같다. 같은 K평균을 사용하기 위한 모델을



만들면서 몇 개의 클러스터가 군집을 잘 나타낼 수 있는지를 평가하는 지표이다. 이를 통해 우리는 SPSS와 달리 3개로 지정하였다. 기존 SPSS와 다르게 나타난 이유는 SPSS에서는 직접 사용자가 지정을 했으며, Expert한 K평균 모델을 사용했기 때문에 다르게 나타난 것으로 보인다. 이를 통해 볼 수 있는 산점도는 다음과 같다.

데이터의 군집 상으로는 X,Y 라는 기준에 따라 분류되었다. 이는 명목형 데이터를 수치상으로 변환했기 때문에 직관성이 떨어진다. 따라서, 구체적 클러스터의 현상을 되짚어 볼 필요가 있다 .



[[1]]

시군명	최종점수	사고년도	사고유형구분	발생건수	사상자수
부천시 :124	Min. :63.48	2012:279	무단횡단 :343	Min. : 1.000	Min. : 1.000
수원시 : 60	1st Qu.:70.28	2013:108	보행노인 : 52	1st Qu.: 4.000	1st Qu.: 4.000
안양시 : 31	Median :73.85	2014: 27	보행어린이 : 69	Median : 4.000	Median : 5.000
의정부시: 31	Mean :73.85	2015: 30	스쿨존내어린이: 18	Mean : 5.285	Mean : 5.608
고양시 : 29	3rd Qu.:78.13	2016: 57	자전거 : 69	3rd Qu.: 6.000	3rd Qu.: 7.000
성남시 : 25	Max. :82.79	2017: 28		Max. :26.000	Max. :26.000
(Other) :251		2018: 22			

중상자수	경상자수	부상자수	거리	사망구분	cluster
Min. : 0.000	Min. : 0.000	Min. :0.0000	Min. :0.01173	FALSE:434	Min. :1
1st Qu.: 1.000	1st Qu.: 1.000	1st Qu.:0.0000	1st Qu.:0.26580	TRUE :117	1st Qu.:1
Median : 3.000	Median : 2.000	Median :0.0000	Median :0.43837		Median :1
Mean : 2.924	Mean : 2.156	Mean :0.2686	Mean :0.52283		Mean :1
3rd Qu.: 4.000	3rd Qu.: 3.000	3rd Qu.:0.0000	3rd Qu.:0.67776		3rd Qu.:1
Max. :22.000	Max. :10.000	Max. :3.0000	Max. :4.50197		Max. :1

```
[[2]]
   시군명        최종점수        사고년도          사고유형구분      발생건수          사상자수
 안산시 :153   Min.   :63.48   2012: 24   무단횡단      : 38   Min.   : 1.000   Min.   : 1.000
 수원시 :105   1st Qu.:68.23   2013: 63   보행노인      : 26   1st Qu.: 4.000   1st Qu.: 4.000
 고양시 : 60   Median :68.23   2014: 58   보행어린이    : 58   Median : 5.000   Median : 5.000
 안양시 : 33   Mean   :70.90   2015:184   스쿨존내어린이: 22   Mean   : 4.965   Mean   : 5.215
 부천시 : 25   3rd Qu.:74.81   2016: 86   자전거        :420   3rd Qu.: 6.000   3rd Qu.: 6.000
 시흥시 : 19   Max.   :82.79   2017: 74                        Max.   :20.000   Max.   :20.000
 (Other):169                   2018: 75
   중상자수        경상자수        부상자수          거리          사망구분      cluster
 Min.   : 0.000   Min.   :0.000   Min.   :0.000   Min.   :0.01197   FALSE:503   Min.   :2
 1st Qu.: 1.000   1st Qu.:2.000   1st Qu.:0.000   1st Qu.:0.34151   TRUE : 61   1st Qu.:2
 Median : 2.000   Median :3.000   Median :0.000   Median :0.53626               Median :2
 Mean   : 1.874   Mean   :2.706   Mean   :0.516   Mean   :0.61017               Mean   :2
 3rd Qu.: 3.000   3rd Qu.:4.000   3rd Qu.:1.000   3rd Qu.:0.77163               3rd Qu.:2
 Max.   :12.000   Max.   :8.000   Max.   :4.000   Max.   :5.34458               Max.   :2
```

```
[[3]]
   시군명        최종점수        사고년도          사고유형구분      발생건수          사상자수
 성남시 :113   Min.   :63.48   2012: 48   무단횡단      : 19   Min.   : 1.000   Min.   : 1.000
 수원시 : 81   1st Qu.:68.23   2013: 81   보행노인      :527   1st Qu.: 3.000   1st Qu.: 3.000
 고양시 : 44   Median :70.81   2014:164   보행어린이    : 52   Median : 4.000   Median : 4.000
 안양시 : 41   Mean   :72.60   2015: 73   스쿨존내어린이: 11   Mean   : 4.104   Mean   : 4.315
 용인시 : 35   3rd Qu.:78.13   2016: 93   자전거        : 35   3rd Qu.: 5.000   3rd Qu.: 5.000
 의정부시: 32   Max.   :82.60   2017: 88                        Max.   :17.000   Max.   :17.000
 (Other) :298                  2018: 97
   중상자수        경상자수        부상자수          거리          사망구분      cluster
 Min.   :0.000   Min.   : 0.000   Min.   :0.0000   Min.   :0.007155   FALSE:482   Min.   :3
 1st Qu.:1.000   1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:0.265478   TRUE :162   1st Qu.:3
 Median :2.000   Median : 1.000   Median :0.0000   Median :0.445798               Median :3
 Mean   :2.457   Mean   : 1.328   Mean   :0.2391   Mean   :0.515413               Mean   :3
 3rd Qu.:3.000   3rd Qu.: 2.000   3rd Qu.:0.0000   3rd Qu.:0.654635               3rd Qu.:3
 Max.   :9.000   Max.   :10.000   Max.   :3.0000   Max.   :4.193380               Max.   :3
```

다음의 총 3개의 클러스트를 통해 나타난 결과이다. 군집 1의 경우 사망사건이 어느 정도 유의하게 나타나는데 이는 무단횡단의 문제로 보여 진다. 군집 2의 경우 대다수가 생존사건인데 이는 자전거사고가 대다수이며, 자전거 사고는 사망으로 직접적으로 안 이어지는 것을 확인할 수 있다. 군집 3은 사망구분 별 대략 30%로 사망확률이 높은 군집이다. 이는 보행노인의 사고 유형의 대부분을 차지하는데, SPSS의 결과와 동일하게 교통사고 발생 시 사망과 관련이 높은 사고는 보행노인사고이며, 사고년도의 분포를 보아 사고의 방지가 필요한 상태로 보인다. 해당 상세 결과로 보아 전반적으로 군집 2는 산점도상으로 군집 1과3의 거리가 차이가 있는 상태이며, 이는 사망구분으로 이어진다고 볼 수 있다.

5. 결론

연구를 통해 교통사고를 겪는 자체가 큰 고통이며 사망사고로 연결되었을 때는 주변 사람들과 가족들에게는 엄청난 비극일이다. 교통사고가 발생하였을 때 긴급조치와 신속한 대응 역시 중요하겠지만 가장 좋은 방법은 교통사고가 발생하지 않는 것이라고 생각한다. 그래서 각 지역에서 교통사고를 줄이기 위해 노력 하고 있지만 우리는 지역별로 교통사고의 발생 건수와 그 원인을 분석하여서 교통사고를 더 많이 줄일 수 있는 효과적이고 효율적인 방법을 제시하고 싶었다.

이 주제를 가지고 시작할 때 교통사고로 인한 사망사고는 대부분은 자동차끼리의 사고, 운전자들의 실수에서 비롯될 것이라고 생각하였다. 하지만, 교통사고로 인한 사망률은 보행노인이 제일 많은 것을 보며 운전자의 경각심도 필수적인 요소이지만 보행자 스스로가 조심해야 한다는 생각이 든다.

아직은 지역마다 발생하는 교통사고의 요인들을 살펴보고 비교하는 정도이지만, 나아가 정확하게 교통사고가 발생할 곳을 특정할 수 있으며 예방할 수 있도록 노력할 것이다.

참고문헌

1. *http://catalk.kr/cars/oecd-road-deaths.html*
2. *http://www.jeollailbo.com/news/articleView.html?idxno=582107*
3. *https://koti10.blog.me/221298478345*
4. *https://blog.naver.com/infoswadcom/221525447673*
5. *http://www.newsis.com/view/?id=NISX20191205_0000851094&cID=10201&pID=10200*