대한지리학회지 제39권 제3호 2004 (457~472)

GIS와 공간 데이터마이닝을 이용한 교통사고의 공간적 패턴 분석

- 서울시 강남구를 사례로 -

이 건 학*

A Study on Spatial Patterns of Traffic Accidents using GIS and Spatial Data Mining Methods: A Case Study of *Kangnam-gu*, Seoul

Gunhak Lee*

요약 : 본 연구의 목적은 GIS와 공간 데이터마이닝 방법을 이용하여 교통사고의 공간적 패턴을 살펴보고, 이웃한 공간 객체와의 공간적 연관성을 탐색하는 것이다. 이를 위하여 서울시 강남구 교통사고 데이터를 이용하여 공간적 경향 분석, 군집 분석 및 군집의 특성 기술, 이웃한 공간 객체와의 연관 분석을 실시하였다. 그 결과, 강남구의 교통사고는 특징적인 4개의 군집 유형을 통해 분류될 수 있으며, 각 군집별로 차별적인 특성들을 보여주고 있다. 또한, 교통사고의 발생 위치와 이웃한 공간 객체들과의 연관성에서는 공간 객체들의 개념수준이나 공간적 관계의 수준에 따라 다양한 규칙들이 발견되었다. 이러한 규칙들은 모두가 유의미하거나 흥미로울 수는 없지만, 맥락에 따라 다양하게 해석될 수 있으며, 보다 심화된 연구를 위한 새로운 가설들로 사용될 수 있을 것이다.

주요어 : 공간 데이터마이닝, GIS, 교통사고, 공간적 관계

Abstract : The purpose of this study is to analyze spatial patterns of traffic accidents and to investigate spatial relations among neighboring spatial objects by applying GIS and spatial data mining methods. This study investigated traffic accident data in *Kangnam-gu, Seoul*, as a case study. As a result, four clusters were emerged based on individual attributes of traffic accidents. Each cluster showed distinctive properties. In spatial associations between individual attributes of traffic accidents and neighboring spatial objects, there were many rules according to concept hierarchy and definition of spatial relations. Although all rules were not be interesting and significant, they could be a clue to investigate more.

Key Words : spatial data mining, GIS, traffic accidents, spatial relations

1. 서 론

1) 문제 제기 및 연구 목적

급격한 경제 성장은 성장과 발달이라는 긍정적 결과와 동시에 사회적 분화, 계층간 대립, 범죄, 교통문제, 환경오염, 도덕적 위기 등과 같은 많은 사회 문제를 야기했다. 그 중에서 교통문제는 자동차의 보급이 급속하게 이루어지면서 가장 시급히 해결해야 할 당면 과제이다. 특히 해마다 수십 만 건씩 발생하고 있는 교통사고는 인적, 경제적 손실이 매우 큰 사회 문제이다.

이러한 교통사고의 예방과 관리를 위해서 많은 연구가 있어 왔고, 많은 대안들이 제시되고 있지만, 전반적인 해결책이 되지 못하고 있는 실정이다. 기존의 여러 가지 연구들은 교통사고의 단순집계나 사고를 유발하는 요인에 대한 분석에만 국한되고 있다. 또한 사고 요인에 대한 분석 역시, 인적, 차량적, 도로환경적 요인들을 복합적으로 고려하지 못하고 있다. 교통사고의 예방과 관리를 위해

* 서울대학교 대학원 졸업(M.A., Graduate school of Seoul National University) guns503@yahoo.co.kr

서는 사고 요인들에 대한 분석과 더불어 교통사고의 특성을 고려한 연구가 진행될 필요가 있다. 교통사고는 이동(movement)이 전제되는 현상이며, 이동은 토지이용, 거주지 패턴, 인구 밀도, 거리 기하구조, 쇼핑 센터 위치, 보건소 위치 등의 공간 객체의 특성에 의존하게 된다. 따라서 본 연구에서는 교통사고의 공간적 특성을 고려한 분석을 수행하고자 한다.

매일, 매시간 발생하고 있는 교통사고 데이터는 사고 특성뿐 아니라 기하학적, 위상적 특성도 포함하는 대용량 공간 데이터이다. 이러한 대용량 교통사고 데이터는 시급한 분석과 해석을 필요로 하기 때문에 기존의 일반적인 통계적 기법이나 공학적 기법으로는 효율적인 분석이 어렵다. 따라서 대용량 공간 데이터베이스에서 효율적인 분석을 수행할 수 있는 새로운 분석 방법론이 필요한데, 본 연구에서는 그러한 분석 방법으로 공간 데이터마이닝을 이용하고자 한다.

공간 데이터마이닝은 급격하게 증가하는 데이터들을 보다 효과적으로 분석하여 유용하고, 의미있는 정보나 지식을 찾기 위해 수행하는 데이터마이닝의 공간적 확장이라고 볼 수 있다. 일반적인 데이터마이닝은 거대한 데이터베이스에서 효율적인 분석을 수행할 수 있는 방법론이지만 공간 데이터에 적용될 경우, 공간 데이터의 특수성으로 인해 그 효과를 기대하기가 어렵다. 공간 데이터마이닝은 원격탐사, 모니터링 시스템, GIS, GPS 등 다양한 기술들의 발달로 공간 데이터의 양이 급격하게 증가하고 있는 환경에서 의미있는 정보와 지식을 추출할 수 있는 적절한 분석 방법이다.

본 연구의 목적은 서울시 교통사고 데이터를 사례로 공간 데이터마이닝 방법을 이용하여 교통사고의 공간적 패턴을 살펴보고, 이웃한 공간 객체와의 공간적 연관성을 탐색하는 것이다.

2) 연구 지역 및 연구 방법

본 연구의 사례 지역은 서울시 강남구이며, 2001년 서울시 교통사고 데이터를 이용하였다. 서울시는 연간 교통사고 건수가 4만5천 여건(2001년)에 달하고, 특히 강남구에서 발생한 교통사고는 3천여건(2001년)으로 서울시에서 가장 많은 사고 발생을 보여주고 있다. 따라서 서울시 강남구 교통사고 데이터는 대용량의 공간 데이터베이스로 공간 데이터마이닝을 적용하기에 적절한 사례가 된다. 본 연구에서는 먼저 우리나라 교통사고 분석 방법의 현황을 살펴보고, GIS와 공간 데이터마이닝을 이용하여 서울시 강남구 교통사고의 공간적 패턴과 공간 객체와의 공간적 연관성을 분석하였다.

본 연구에서 사용된 교통사고 데이터는 사고 발생시 담당 경찰관에 의해서 기재되는 교통사고 통계원표와 경찰청, 도로교통안전관리공단에서 발행하는 각종 통계책자에 기재된 통계 자료이다. 교통사고 데이터는 사고 발생시 각 관할 경찰서의 담당 경찰관에 의해서 일차적으로 데이터가 구축되는데 이러한 데이터를 토대로 중앙 전산시스템에 데이터베이스가 구축된다. 하지만 데이터 구축에서 여러 가지 원인[1]으로 인하여 데이터의 무결성이 보장되지 않기 때문에 분석의 정확성이 떨어진다. 그러나 데이터마이닝은 데이터의 결측치와 에러에 대한 유연함을 가지고 있기 때문에 이러한 어려움은 어느 정도 극복될 수 있다.

본 연구에서 사용된 소프트웨어는 ESRI社의 ArcGIS 8.2, ArcView 3.2, National Institute of Justice(NIJ)의 CrimeStat 1.0을 사용하였고, 교통사고 데이터의 마이닝을 위해서 SPSS社의 Clementine 6.5, Insightful社의 Insightful Miner를 이용하였고, 분포 패턴의 특성 기술을 위해서 Insightful社의 S-Plus for windows를 이용하였다.

2. 공간 데이터마이닝의 정의와 기법

공간 데이터마이닝이란 데이터마이닝의 공간적 확장이다. 공간 데이터가 가지는 특수성을 고려한 마이닝이라고 볼 수 있다. 공간 데이터마이닝의 정의는 학자마다 조금씩 다르지만, Koperski *et al.*(1998)은 "함축적인 지식, 공간적 관계, 또는 공간 데이터베이스에서 명시적으로 저장되어 있지 않은 패턴들의 추출"로 정의하고 있으며, Miller and Han(2001, 16)는 "공간 데이터마이닝을 지리적 공간에 분포하는 객체, 이밴트의 분포에서 흥미로운 패턴들을 추출하기 위한 컴퓨터적 툴의 응용"이라고 했다. 여기에서 패턴들은 일반적인 데이터마이닝에서의 비공간적 속성의 흥미로운 패턴뿐

아니라 개별적 객체와 이벤트의 공간적 속성(객체의 모양, 범위), 객체와 이벤트들 사이의 시·공간적 관계에 대한 패턴들을 포함한다. 따라서 공간 데이터마이닝은 교통사고 데이터에서 교통사고의 공간적 속성과 시·공간적 관계에 대한 패턴을 찾을 수 있으며 축약된 방법으로 교통사고 데이터의 규칙성을 나타낼 수 있다.

공간 데이터마이닝의 주요 기법은 공간적 군집 분석[2], 공간적 분류화[3], 공간적 연관 분석[4], 공간적 이례 분석[5], 공간적 요약화[6] 등이 있다. 공간 데이터마이닝은 전술한 것처럼 공간 데이터의 특수성을 고려한 마이닝 알고리즘을 포함하고 있다. 공간 데이터베이스 측면에서 공간 데이터의 특수성은 공간 데이터마이닝을 위한 알고리즘의 선택과 수행에 기초가 되는 공간 객체들 사이의 공간적 관계에 대한 정의와 공간 데이터베이스의 효율적 질의들을 통해서 고려되고 있다. 데이터베이스에서 공간적 관계는 조작적 정의를 통해서 관계형 데이터베이스 테이블에 저장되고, 질의를 통해서 공간 데이터마이닝 작업이 수행된다.

본 연구에서는 교통사고의 공간적 경향을 살펴보기 위하여 공간적 군집 분석을 수행했고, 일반적인 군집 분석을 통해 교통사고의 특성을 몇 개의 유형으로 분류하였다. 그리고 공간적 연관 분석을 통해 교통사고와 이웃한 공간 객체와의 연관성을 살펴보았다.

3. 우리나라 교통사고 분석 방법의 현황

우리나라의 자동차 등록현황을 보면 1970년 12만 8천 여대에서 2001년 현재 1290만 여대로 100배 이상 증가했다. 자동차의 보유의 보편화는 우리나라의 급격한 경제성장에 따른 발달을 보여줌과 동시에 교통사고라는 역기능을 가져왔다. 국내에서 이루어지고 있는 교통사고 관련 연구 동향을 살펴보면, 크게 거시적 관점과 미시적 관점에서 고찰할 수 있다. 거시적 관점에서는 전국 또는 지역별로 전체적인 동향이나 주요 항목을 중심으로 하는 분석을 말한다(오재학·이대근, 1995, 17). 즉 교통사고 데이터베이스에서 항목별 집계에 의한 집계분석이라고 볼 수 있다. 교통사고 데이터베이스는 최초 사고 발생 후 관련자의 신고로 경찰관이 24시간 내에 통계원표 및 보충표를 작성하여 구축된다. 이렇게 작성된 데이터는 각 관할 경찰서 별로 전산 입력되고, 관할 경찰서에서 수집된 데이터들은 경찰청 본청의 전산실 서버의 데이터베이스로 구축된다. 이 데이터베이스를 토대로 분석이 이루어지며 분석된 결과는 매년 책자로 발간된다.[7] 이러한 통계분석 결과는 교통사고의 전반적인 경향 파악에 도움이 되지만, 특정한 지점 내지 지역에 대한 사고 대책을 수립하는데 직접적인 도움을 주지 못하기 때문에 특정한 지점에 대한 분석을 수행하기 위해서는 미시적인 교통사고 분석을 도로별로 수행해야 한다.

한편 미시적인 관점에서는 전체적인 사고 경향보다 교통사고 요인에 대한 특성을 밝히고자 하는 분석이 이루어지는데, 크게 교통사고 요인에 대한 특정화를 위한 방법과 잠재적인 사고위험성을 예측하기 위한 연구로 이루어지고 있다. 전자는 교통사고에 대한 발생 요인을 여러 가지 분석을 통해 유형화를 시키고, 후자는 교통량, 주행속도, 사고유형, 도로환경 등의 다양한 요인이 교통사고 건수에 미치는 영향에 대한 모형을 만들고 교통사고 위험도 평가를 수행한다. 이러한 분석들은 목적이 서로 다르지만 교통사고의 발생 특성이나 사고 요인 분석이라는 점이 공통된다. 교통사고의 사고 요인이 되는 발생 특성은 크게 인적 요인, 차량적 요인, 도로환경적 요인으로 구분하여 분석하고 있지만, 요인들이 독립적으로 작용한 사고보다 복합적으로 발생한 사고가 많다. 따라서 교통사고의 정밀한 분석은 복합적 요인에 대한 체계적인 분석을 요하지만, 대부분의 연구는 일부의 요인에 주목하여 개별적인 접근을 시도하고 있다. 김효종, 서채연(1995)은 선형 회귀분석을 통해서 교통량과 교통량과 교통사고와의 관계를 분석했고, 이주형 외(1990)는 교통사고 발생 특성, 지역과 도로 형태에 따른 사고요인에 대한 분석을 실시했고, 고상선(1996)은 사고자의 인적인 요인에 초점을 맞춘 분석을 수행했다.

교통사고는 두 개 이상의 이동하는 객체나 고정된 객체와 이동하는 객체 사이의 원하지 않는 상호작용의 산물이다. 또한 이동은 보행자든 운전자든 토지이용시스템, 거주지 패턴, 인구 밀도, 거리 기하구조, 작업장, 쇼핑센터, 보건소 위치 등 여타 다른

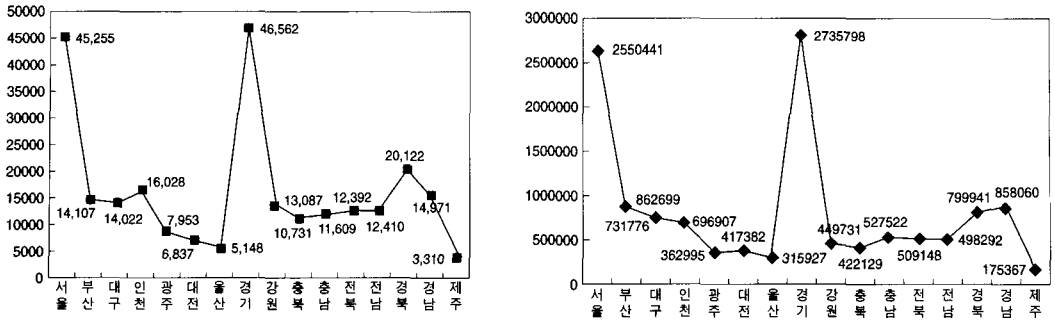

그림 1. 전국 교통사고 발생 건수와 자동차 등록대수(2001년)

출처: 도로교통안전관리공단, 2002, 교통사고 통계분석

교통 유발자의 함수일 수 있다. 즉 교통사고는 시공간 이벤트로 보아야 할 것이며, 여러 가지 다른 환경의 조합을 통해서 추정될 수 있다(Whitelegg, 1987, 162). 하지만 현행 교통사고 분석은 대부분 교통사고 자체의 원인규명에 초점을 맞추고, 교통사고의 시공간적 본질을 파악하지 않고 있다. 또한 교통사고의 수치적 해석에만 주목하고 있으며, 공간적 패턴에 대한 관심을 등한시하고 있다.

따라서 본 연구에서는 교통사고 발생의 공간적 패턴에 주목하여, 교통사고의 공간적 패턴과 공간객체와의 공간적 연관성을 살펴보았다.

4. GIS와 공간 데이터마이닝을 이용한 교통사고의 공간적 패턴 분석

1) 서울시 교통사고 현황

서울은 2001년 현재 광역시를 포함한 전국에서 경기도에 이어 가장 높은 교통사고 발생 건수를 나타내고 있다. 또한 자동차 등록대수에 있어서도 경기도에 이어 두 번째로 높은 수치를 보이고 있다(그림 1).

서울시 구별 교통사고 발생과 특성을 살펴보면 그림 2에서처럼 구별 발생 사고 건수는 강남구, 송파구 일대가 가장 많으며, 성동구, 동작구, 은평구, 도봉구, 서대문구가 사고 발생이 가장 적은 지역이었다. 하지만 사망자수와 중상자수, 경상자수를 구별로 살펴보면 발생 빈도와는 조금씩 상이한 결과를 보여준다.

2) 교통사고 데이터베이스

공간 데이터에서 공간 데이터마이닝은 일반적인 마이닝과는 달리 공간 데이터의 특수성을 다룰 수 있는 공간 데이터베이스의 구축을 요구한다. 이 작업은 매우 시간 소모적이고, 많은 주의를 필요로 한다. 교통사고 데이터의 공간 데이터베이스는 크게 두 가지로 구축되었다. 교통사고 위치와 이웃한 공간 객체 데이터(공간 데이터)와 교통사고의 속성 데이터(속성 데이터)로 구축하였다.

수집된 데이터는 원시 데이터이기 때문에 분석에 사용하기 위해서는 몇 가지 데이터 정제 및 전처리 작업이 필요하다. 이러한 데이터 정제 및 전처리를 통해 구축된 공간 데이터베이스 항목은 다음과 같다(표 1).

공간 데이터베이스의 구축은 많은 부분 GIS를

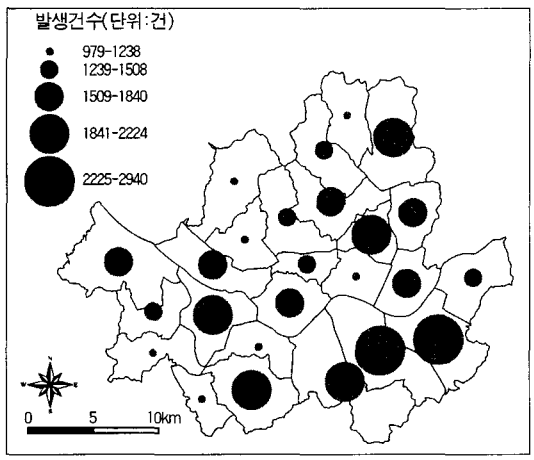

그림 2. 서울시 구별 교통사고 발생 현황(2001년)

표 1. 분석에 이용된 데이터 항목

구 분	데이터 항목	데이터 타입
교통사고 개별속성	사고내용, 사고유형, 발생요일, 발생월, 발생주야, 사고자 성별, 도로형태, 차도폭, 도로선형, 노면상태, 중앙분리표시, 신호등, 속도규제, 토지이용, 날씨, 차량종류, 차량파손정도, 통행목적, 음주운전, 면허경력, 보호장구, 통행형태	명목형
	사망자수, 중상자수, 경상자수, 물적 피해액, 사고자 연령	수치형
이웃한 공간객체	도로시설, 문화교육시설, 산업시설, 서비스시설, 의료후생시설, 주택시설, 행정기관, 지형, 하천	

이용했다. GIS는 공간 데이터를 매우 효율적으로 다루고, 공간 데이터베이스 구축을 위한 여러 가지 기능을 제공했으며, 특히 공간 객체들의 거리와 위상관계에 기반한 공간 질의와 같은 기능은 분석에 필요한 공간 데이터 구축에 결정적인 역할을 하였다.

3) 공간 데이터마이닝 및 패턴 발견

강남구 교통사고의 공간적 패턴은 먼저 강남구 교통사고 발생의 공간적 경향을 파악하고, 강남구 교통사고 속성에 대한 유형화를 통해서 유형별 공간적 분포 패턴을 개관할 수 있다. 교통사고의 유형화는 일반적인 군집 분석 후 군집에 따른 특성화를 통해서 수행했다. 이후 교통사고와 이웃한 공간 객체와의 공간적 연관성을 살펴보기 위해서 군집별로 공간적 연관 분석을 실시했다.

(1) 공간적 경향성

서울시 강남구 교통사고는 2001년 한해 동안 총 3002건의 사고가 발생했다. 발생된 사고는 담당 경찰관에 의해서 통계원표로 작성되는데, 공간적 분포를 파악하기 위해 통계원표 항목 중 X, Y 좌표를 이용하여 공간 데이터로 변환했다. 총 3002건의 사고 중 동일한 좌표[8]를 가진 사례가 많기 때문에 공간 상에 표현되는 사고 발생지점은 그림 3처럼 총 102건의 사고로 표시되었다. 지도상에 표시된 사고 발생지점은 가장 남쪽의 대모산 일대를 제외하고 강남구 전체에 걸쳐 도로 네트워 상에서 골고루 분포하고 있으며, 몇몇 지점들은 공간적 군집을 보여주고 있다. 하지만 발생지점별 사고 건수가 틀리기 때문에 분포상의 모습으로 공간적 군집을 단정지을 수는 없다.

그림 4는 발생지점별로 사고 건수를 나타낸 것

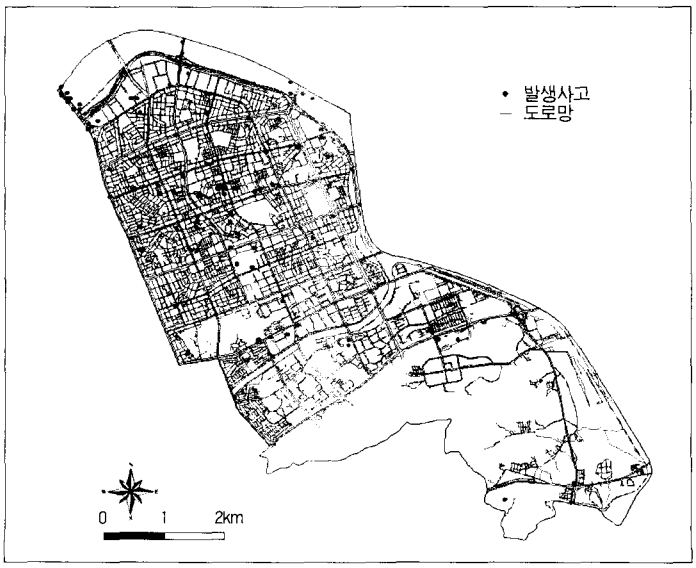

그림 3. 강남구 교통사고 발생 위치도

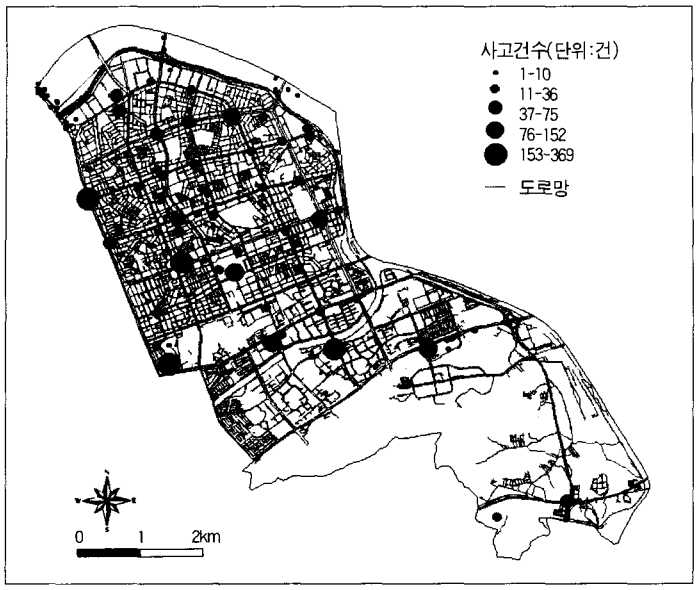

그림 4. 강남구 교통사고 발생 건수

인데, 그림 3과는 달리 지점별로 다른 사고 건수를 보여주고 있다. 그림 4에서 보여지듯이 강남대로에서 논현역, 양재역 부근과 테헤란로와 언주로가 교차하는 역삼역 부근과 삼성로에서 양재대로 방면의 개포 근린공원 일대 부근에서 150건 이상 빈번한 사고가 발생했으며, 역삼로의 개나리 아파트 부근, 영동대로와 테헤란로가 교차하는 삼성역 부근, 양재대로에서 삼성의료원 부근, 삼성로와 압구정로가 만나는 부근, 남부순환로와 선릉로가 교차하는 도곡역 부근에서 역시 70건 이상의 잦은 사고가 발생했다. 한남대교 부근의 사고는 공간적으로 매우 밀집해서 발생했지만, 실제 사고 건수는 적은 곳임을 보여준다.

교통사고 분석에서 교통사고가 빈번하게 발생하

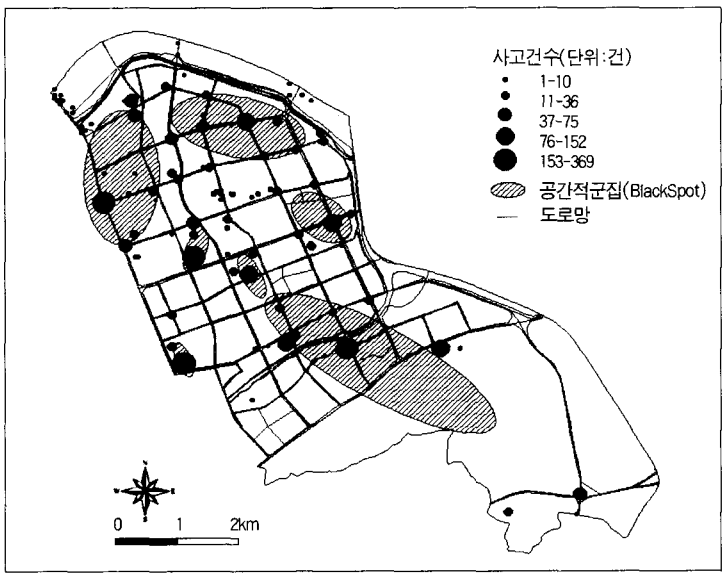

그림 5. 강남구 교통사고 분포의 공간적 군집

는 지점들을 다발지점(black-spot)이라고 부른다. 이것은 동일 장소에서 상대적으로 많은 사고가 발생하는 지점들을 말한다. 공간상에서 같은 위치더라도 많은 사례수를 포함하기 때문에 이러한 지점들은 공간적으로 밀집되어 분포하는 공간적 군집에 많은 영향을 미치게 된다. 그림 5는[9] 강남구 교통사고의 공간적 군집을 보여준다. 공간적 군집은 모두 7군데가 발견되었는데, 대부분 사고가 가장 빈번하게 발생하는 지점들을 포괄하고 있다. 이것은 그 지점의 사례수가 군집을 판별하는데 중심점으로써의 역할을 하고 있다는 것을 의미한다.[10]

2001년 강남구에서 발생한 교통사고의 전체적인 경향은 사고 내용에 있어서 사고의 정도가 미비한 경상사고가 많은 빈도를 차지하고, 차대차의 사고가 주로 일어나며, 계절상 겨울철에 많이 일어나고 있으며, 주야를 구분할 때, 거의 비슷한 비율로 나타나고 있다. 사상자수는 많지 않으며, 물적 피해액은 평균 143만원이다. 사고자의 연령은 주로 30대 후반이 많으며, 대부분의 사고들이 남성 운전자에 의해서 나타나고 있다. 사고의 특성을 살펴보면, 강남구의 통행목적은 업무상의 목적이 가장 높으며, 음주운전으로 인한 사고는 거의 일어나지 않았으며, 또한 특이하게 대부분 면허가 10년 이상으로 운전에 대한 숙련도가 높은 사람이 많은 사고를 야기하는 것으로 나타났다. 이는 운전에 대한 숙련도에 따른 자만감이 사고의 위험성을 높인다고 짐작할 수 있다. 사고가 일어난 도로환경은 단일로에서 주로 발생했고, 차도폭은 13m 이상의 넓은 대로이고, 도로선형은 대부분 직선도로에서 주로 발생했다. 강남구의 도로들이 대부분 직선의 넓은 대로임을 감안할 때, 매우 일반적인 현상으로 해석된다. 사고발생시 신호등은 대부분 점등상태로 정상적인 신호가 작동되고 있는 곳이었음을 알 수 있다. 그리고 도로상의 속도규제는 제한속도가 없는 곳이 대부분이었고, 날씨는 맑은 날, 대부분의 사고가 발생했다. 차량적 특성으로는 주로 자동차에 의해서 사고가 발생했지만 대부분의 경우 차량파손 정도는 거의 없는 경미한 사고임을 알 수 있다.

(2) 군집 분석

강남구 교통사고를 체계적이고 집중적으로 관리하기 위해 몇 개의 유의미한 유형으로 분류하는 것은 매우 중요한 작업이다. 따라서 본 연구에서는 군집 분석을 통해서 강남구 교통사고를 몇 개의 유형으로 분류하였다. 군집 분석은 최초 연구자에 의해 군집의 수를 결정해야 하는데, 이러한 결정은 매우 탐색적이다. 군집 분석의 여러 가지 알고리즘 중에서 Kaufman and Rousseeuw(1990)에 의해서 개발된 PAM(Partitioning Around Medoids)은 군집 분석 결과를 실루엣 그래프를 통해서 제시해 준다. 실루엣 계수는 객체가 분류된 군집에 실제로 얼마나 속하는지를 나타내는 척도로 -1에서 1까지의 범위를 가진다. 1에 가까울수록 그 군집에 속하는 정도가 높다. 따라서 이러한 평균 실루엣 계수의 값이 높은 K의 수가 적절한 군집의 수가 된다. 탐색적으로 K의 개수의 변화에 따라 평균 실루엣 폭을 구하였는데, 4개의 군집이 가장 적절한 것으로 나타났다. 군집 분석은 SPSS社의 Clementine6.5를 이용하여 K-means 알고리즘을 사용하여 분석을 실시했다. Clementine은 특히 범주형 데이터와 수치형 데이터를 모두 다룰 수 있기 때문에 본 연구에서 사용된 데이터 유형에 적합하며, 군집들에 대한 특성화 기술이 매우 용이한 결과물을 제시하기 때문에 매우 유용한 툴이라 볼 수 있다. 강남구의 교통사고 데이터의 군집 분석 결과는 그림 6과 같다.

군집별로 교통사고의 공간적 분포를 살펴보면 같은 위치에서 발생했더라도 사고특성에 따라 다른 군집으로 분류되며, 빈도가 가장 높은 군집의 유형과 발생 위치에 따른 각 군집별 사고 건수는 그림 7과 같다.

군집 1, 2, 4의 유형이 강남구 교통사고의 대부분의 유형을 차지하고 있기 때문에 사고가 잦은 지점 역시 1, 2, 4의 군집이 대다수를 차지하고 있

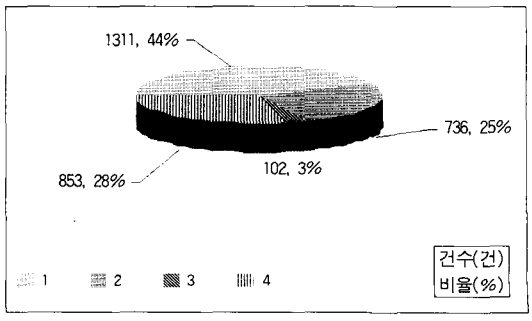

그림 6. 군집 비율

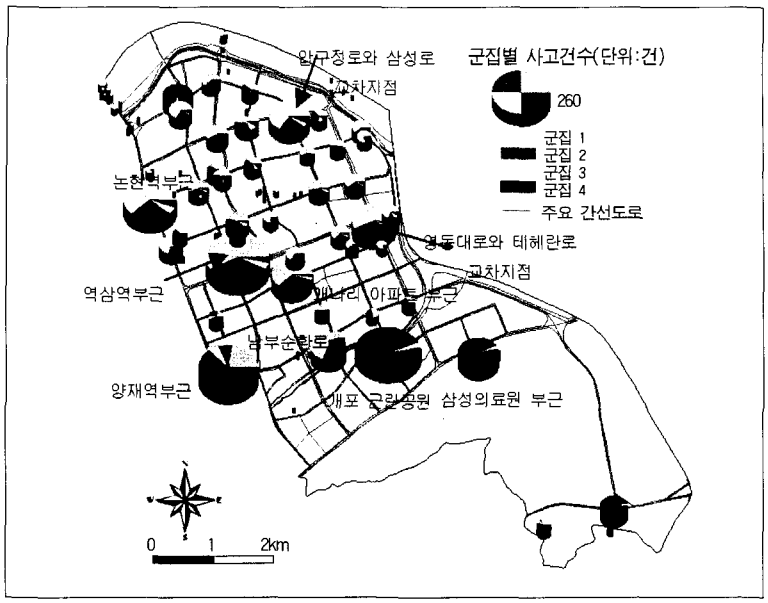

그림 7. 군집별 사고 발생 건수

다. 군집 1은 1,311건으로 가장 높은 비율로, 전체적으로 골고루 분포하며, 특히 강남대로의 논현역 부근, 역삼역 부근의 사고는 매우 빈번한 사고가 발생하는 동시에 군집 1의 유형을 띠고 있다. 또한 압구정로와 삼성로가 교차하는 부근이나, 역삼로 개나리 아파트 부근, 남부순환로 일대 도곡역 부근과 같이 두번째로 빈도가 높은 지점들도 군집 1의 특성을 보이고 있다. 군집 2는 전체 25%를 차지하면서 압구정로 일대와 학동로 일대, 선릉로와 봉은사로 일대, 남부 순환로 일대의 사고유형을 나타낸다. 군집 3은 가장 적은 유형으로 한남대교 일부와 남부 순환로 도곡역 부근의 사고는 적은 빈도지만 군집 3의 특성을 나타내고 있다. 군집 4는 전체 28%로 강남대로 양재역 부근, 개포 근린공원 부근의 높은 빈도의 사고 중 대부분을 차지하는 유형이며, 양재대로 삼성의료원 부근, 세곡동 세곡교 부근의 사고도 군집 4의 유형을 띠고 있다.

(3) 군집의 특성화

각 군집들에 대한 특성화는 Han and Fu(1996)가 제시한 속성 지향 귀납법에 의한 일반화 기반 마이닝을 수행하였다. 속성 지향 귀납법은 데이터들의 개념 계층을 만들고, 이에 따른 일반화를 수행한다. 교통사고의 개별적 속성에 대한 개념 계층은 선행 연구에서 제시한 사고 발생특성의 구분과 사고일반에 관한 속성에 따라 구성하였다. 전체 강남구 교통사고의 경향과 비교해 볼 때, 각 군집들은 사고일반에서 각 군집의 차별적인 특성을 나타내고 있으며, 사고특성에서는 주로 도로환경적 요인이 군집들의 차이를 보여주고 있다.

전체적인 교통사고의 경향과 유사한 특성은 배제하고 각 군집별 특성을 기술하면 다음과 같다.

군집 1은 주로 사고 내용이 중상사고의 경우가 높으며, 사고 요일은 전일에 걸쳐 비슷한 비율로 발생했지만 상대적으로 수요일에 가장 많은 빈도를 보였으며, 전체적인 경향과 달리 야간 교통사고율이 높은 특성을 보였다. 사고시 물적 피해액 역시 전체 평균 143만원 보다 높은 160만원의 사고로 중상사고의 특성을 뒷받침한다. 나머지 사고특성은 신호등이 없는 곳에서 발생한 특성을 제외하고는 강남구의 전체적인 경향과 매우 유사한 형태를 보였다.

군집 2는 전체적인 경향과 유사한 형태를 나타내는 사고 유형이지만 사고자의 연령이 전체 평균 연령보다 약간 낮으며, 군집 1처럼 신호등이 없는 곳에서 주로 발생한 사고 유형이다.

군집 3은 발생일시에서 다른 군집이 12, 1월과 같은 겨울에 발생한 것과 달리 상대적으로 5월, 즉

봄에 주로 발생했으며, 월요일에 많이 발생한 사고 유형이다. 군집 1과 마찬가지로 야간사고율이 상대적으로 약간 높으며, 사고시 물적 피해액은 군집들 중 가장 낮은 비율을 보이고 있으며, 사고자의 연령은 평균 40세로 전체 평균보다 약간 높은 고령의 운전자에 의한 사고 유형이었다. 사고자의 성별에서도 다른 군집과 달리 상대적으로 전체 평균 83%에 비해서 남자 운전자의 비율이 76%로 낮았다.

군집 4는 발생일시에서 다른 군집들이 월, 화, 수요일 같은 주초에 주로 발생하는 반면, 금요일과 같은 주말에 상대적으로 많이 발생한 유형이다. 또한 통행목적에 있어서 개인용무 중 사고가 주로 발생했으며, 도로환경적 특성에서 다른 군집과는 달리 교차로상에서 일어난 사고유형이었고, 차도폭 역시 3m 이상의 좁은 차로에서 사고가 발생했다. 그리고 사고발생 장소의 토지이용은 대부분이 시가지에서 발생한 것과 달리 비시가지 지역의 사고 유형을 나타냈다.

(4) 공간적 연관 분석

연관 분석은 대용량 데이터에서 자주 함께 나타나는 속성값의 조건이나 규칙들을 찾아 속성들 사이의 흥미로운 관계를 보여준다. 연관 분석을 통한 결과들은 명시적으로 나타나지 않지만 유의미한 패턴들이나 규칙들이며 현재 진행중인 연구나 향후 연구에 좋은 실마리를 제공할 수 있다. 본 연구의 분석에서는 교통사고의 비공간적인 속성들 사이의 연관 규칙과 이웃한 공간 객체와의 연관 규칙을 탐색했다. 연관 분석을 위한 소프트웨어는 SPSS社 Clementine의 apriori 모델을 사용하였으며, 앞서 구축된 공간 데이터베이스를 토대로 분석을 실시하였다. 이웃한 공간 객체와의 연관 분석은 선행적으로 공간 객체의 선정과 공간적 관계 정의가 필요하다. 공간 객체는 일반적인 지형도의 객체들을 선정하였고, 공간적 관계는 공간 객체들 사이의 거리에 기반한 위상적 관계만을 설정했다. 세부적인 분석은 크게 두 가지 형태로 진행되었다. 첫째, Han *et al.*(1999)이 제시했던 다수준 연관 규칙의 도출을 위해 공간 객체들의 개념 계층을 구성하여 추상화 수준에 따른 연관 규칙을 찾았다. 둘째, Koperski *et al.*(1995)의 공간적 관계 정의를 이용하여 위상적 계층[11]에 따른 연관 규칙을 찾았다. 공간적 관계를 지시하는 공간적 진술은 2수준[12]까지 분석하였다.

분석을 위한 개념 계층은 영역 전문가에 의해서 주어지거나 연구자에 의해서 생성할 수 있는데, 본 연구에서는 일반적인 지형도의 공간 객체를 선정했기 때문에, 수치지형도의 레이어의 분류코드에 의거하여 다음 표 2와 같은 개념 계층을 구성하였다.

표 2. 공간 객체의 개념 계층

대분류	수준1	수준2	대분류	수준1	수준2
도로환경	도로시설	교차부	인공물	서비스	금융시설
		교량			숙박시설
		보행시설			운수시설
		편의시설		의료후생	병원
인공물	문화교육	교육시설			사회복지
		언론시설			*아동복지*
		종교시설		*주택*	*주택*
		체육시설		행정기관	지방행정
	산업	공업시설			치안행정
		농업시설			정부투자
		상업시설			기타행정
자연환경	*산*	*산*			
	하천	*하천*			

표 3. 공간적 관계의 개념 계층

일반화된 진술 (coarse predicate)	상세화된 진술 (refined predicate)
close_to	intersects, adjacents, nearby

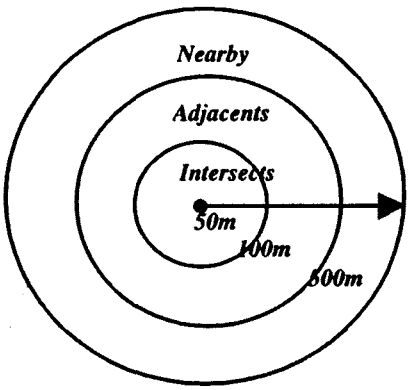

그림 8. 공간적 관계 정의

밑줄이 그어진 객체는 각각의 수준에서 계층 분류가 어려운 것들로 이전의 상위 수준의 것을 그대로 사용하였다. 개념수준1은 개념수준2 보다 추상화가 높은 객체들로 구성된다. 이웃한 공간 객체들의 위상적 계층은 2수준으로 구성하였는데, 크게 일반화된 진술과 보다 상세화된 진술로 구분하였다(표 3).[13]

공간 객체들의 이웃에 대한 공간적 관계 정의는 거리에 기반하였다. 모든 교통사고 발생 위치의 점 객체들 사이의 거리를 계산한 결과, 평균거리가 2255m였고, 평균거리를 기준으로 두 점 어디에도 인접하지 않는 버퍼거리를 1000m로 두었다. 버퍼거리를 제외한 500m 이내에 있는 객체들을 교통사고와 인접한 공간 객체로 간주했다. 따라서 일반화된 진술에 따른 이웃에 대한 정의는 교통사고로부터 500m 이내의 객체들로 정해졌다.

또한 일반화된 진술의 거리를 그림 8과 같이 세 가지의 상세화된 진술 형태로 세분하였다. 50m 이렇게 정의된 공간적 관계는 객체들의 공간적 위치를 기반으로 하여 참, 거짓의 이분형태로 별도의 테이블에 저장했다.

연관 분석에서 매우 중요한 부분은 최소 지지도와 최소 신뢰도를 지정해 주는 것이다. 지지도는 조건이 되는 속성이 가지는 확률을 말하는데, 위상적 계층의 수준에 따라 달리 적용하였다. 일반화된 수준에서는 최소 지지도는 모든 개념수준에 상관없이 50% 이상의 지지도를 보이는 것으로 제한했으며, 상세화된 수준에서는 50%의 지지도를 보다 낮은 16.7%의 비율로 정하였다. 이것은 상세화된 위상계층의 객체 수 총합이 일반화된 수준의 객체수와 같으므로 전체 데이터 셋의 객체 수에서 50% 임계치를 만족시키는 규칙이 도출되기 어렵기 때문이다. 신뢰도는 이러한 조건이 되는 속성이 결과가 되는 속성과 함께 존재할 확률을 말하는데, 각 군집의 사례수가 다르므로 최소 지지도를 만족시키더라도 최소 신뢰도를 만족시킬 수 있는 확률이 달라질 수 있기 때문에 군집의 비율에 따라 최소 50% 이상의 신뢰도를 만족시키도록 적용시켰다. 또한 개념수준에 따라 보다 높은 수준은 더 많은 사례를 포함하기 때문에 수준이 높을수록 보다 높은 신뢰도를 적용시켰다. 각 군집에 대한 최소 신뢰도는 다음 표 4와 같다.

교통사고 속성들의 연관성은 개념 계층을 단일한 수준에서 분석을 실시했고, 최소 지지도 50%를 각 군집마다 동일하게 적용했으며, 최소 신뢰도는 공간적 연관 분석의 개념수준1과 동일하게 각 군집마다 비율수준의 70%를 적용시켰다. 교통사고 개별 속성들의 연관 분석 결과는 표 5와 같다.

도출되는 연관 규칙들은 매우 많았지만 신뢰도 X 지지도별로 정리하여, 상위 10개의 규칙들만 추출했고, 군집마다 흥미로운 규칙을 찾기 위해서 2개 이상의 군집에 공통된 규칙은 일반적인 규칙으로 간주하여 배제시켰다. 신뢰도는 발견된 규칙의 정확성을 의미하고, 지지도는 충분히 큰 빈도를 의미하기 때문에 이 두 가지 모두를 만족시키는 규칙을 찾기 위해 신뢰도와 지지도를 곱한 값이 큰 규칙들을 추출했다.

위의 결과에서 군집 1은 개별 속성과의 연관된

표 4 군집별 최소 신뢰도

군집유형	비율(%)	수준1(비율의 70%)	수준2(비율의 60%)
1	44	30.8	20.4
2	25	17.5	15
3	3	2.1	1.8
4	28	19.6	16.8

표 5. 교통사고 속성들의 연관 규칙

조건(IF)	규칙수	결과(Then)
규칙 없음	0	군집1
중앙분리표시==서비스구역 & 속도규제==규제없음 & 사망자수==없음 (2971:99.0%, 0.246) 속도규제==규제없음 & 사망자수==없음 (2971:99.0%, 0.246) 중앙분리표시==서비스구역 & 사망자수==없음 (2971:99.0%, 0.246)	3	군집2
노면상태==포장 & 중앙분리표시==서비스구역 & 속도규제==규제없음 (2956:98.5%, 0.035) 노면상태==포장 & 사망자수==없음 (2955:98.4%, 0.035) 경상자수==적음 & 노면상태==포장 & 속도규제==규제없음 (2955:98.4%, 0.035) 노면상태==포장 & 중앙분리표시==서비스구역 & 사망자수==없음 (2955:98.4%, 0.035) 노면상태==포장 & 속도규제==규제없음 & 사망자수==없음 (2955:98.4%, 0.035) 경상자수==적음 & 노면상태==포장 & 사망자수==없음 (2953:98.4%, 0.035)	6	군집3
도로선형==직선 (2889:96.2%, 0.292) 노면상태==포장 & 도로선형==직선 (2879:95.9%, 0.293) 경상자수==적음 & 도로선형==직선 (2887:96.2%, 0.292) 노면상태==포장 & 중상자수==적음 (2966:98.8%, 0.284)	4	군집4

규칙을 발견할 수 없었으며, 군집 2는 중앙분리표시와 속도규제와 같은 도로환경적 요인이 강하게 연관된 규칙이 발견되었다. 사망자 수는 강남구 교통사고의 전체적인 경향이 매우 적게 발생하기 때문에 의미를 가지지 못한다. 군집 3은 노면상태와 속도규제와 같은 도로환경적 요인이 강하게 연관되어 있고, 경상자수와 같은 인적인 요인과도 연관된 규칙이 발견되었다. 군집 4는 노면상태와 도로선형과 같은 도로환경적 요인이 강하게 연관되어 있다. 전체적으로 도로환경적 특성들이 군집의 유형에 큰 영향을 미치고 있음을 알 수 있다.

교통사고와 공간 객체와의 공간적 연관성에서 발견된 규칙들은 다음과 같다. 신뢰도 X 지지도별로 정리했으며, 속성들의 연관처럼 2개 이상의 군집에 공통된 규칙은 일반적인 규칙으로 간주하여 배제시킨 후 도출되는 규칙만 아래 표 6에 정리했다.

표 6은 일반화된 공간적 진술에서 개념수준1의 공간 객체와의 연관성을 보여주는 것으로 군집 1과 강하게 연관된 규칙은 발견되지 않았으며, 군집 2는 서비스와 문화교육, 의료후생과 같은 인공물들이 함께 이웃해 있으면 그 특성을 나타낸다. 군집 3은 의료후생, 서비스, 주택, 문화교육과 같은 인공물들이 강하게 연관되어 있고, 도로시설과도 연관성을 보이고 있다. 군집 4는 문화교육 시설이 공통적으로 있으며, 동시에 서비스, 도로시설, 의료후생과 같은 인공물들이 함께 이웃해 있으면 높은 연관성을 나타낸다.

표 6. 일반화된 공간적 진술-개념수준1

조건(IF)	규칙수	결과(Then)
규칙 없음	0	군집1
close_to_서비스 & close_to_문화교육 & close_to_의료후생 (2949:98.2%, 0.248)	1	군집2
close_to_도로시설 & close_to_의료후생 & close_to_주택 (2898:96.5%, 0.035) close_to_의료후생 & close_to_주택 (2898:96.5%, 0.035) close_to_문화교육 & close_to_의료후생 & close_to_주택 (2897:96.5%, 0.035)	3	군집3
close_to_도로시설 & close_to_서비스 & close_to_문화교육 (2951:98.3%, 0.286) close_to_서비스 & close_to_문화교육 (2951:98.3%, 0.286) close_to_도로시설 & close_to_문화교육 & close_to_의료후생 (2950:98.3%, 0.286) close_to_문화교육 & close_to_의료후생 (2950:98.3%, 0.286)	4	군집4

표 7. 일반화된 공간적 진술-개념수준2

조건(IF)	규칙수	결과(Then)
규칙 없음	0	군집1
규칙 없음	0	군집2
close_to_교육 & close_to_주택 (2886:96.1%, 0.035)	1	군집3
close_to_보행 & close_to_교육 & close_to_금융 (2914:97.1%, 0.287)	1	군집4

표 8. 상세화된 공간적 진술-개념수준1

조건(IF)	규칙수	결과(Then)
규칙 없음	0	군집1
intersects_도로시설 & nearby_산업 (1191:39.7%, 0.299), nearby_산 (1401:46.7%, 0.247)	2	군집2
규칙 없음	0	군집3
규칙 없음	0	군집4

표 9. 상세화된 공간적 진술-개념수준2

조건(IF)	규칙수	결과(Then)
규칙 없음	0	군집1
규칙 없음	0	군집2
nearby_주택 (1434:47.8%, 0.041)	1	군집3
nearby_종교 & nearby_사회복지 (1765:58.8%, 0.339) nearby_지방행정 (1311:43.7%, 0.358) nearby_아동복지 (1664:55.4%, 0.278)	3	군집4

표 7은 개념수준2의 공간 객체와의 연관성을 나타낸 것인데, 표 6처럼 군집 1에서는 어떤 연관도 발견할 수 없었다. 군집 2는 표 6에서는 연관성을 보이는 규칙이 도출되었지만 보다 세부적인 객체와의 연관성은 나타나지 않았고, 군집 3은 표 6의 의료후생, 서비스, 주택 같은 인공물 중에서 특히 교육시설이나 주택이 강하게 연관되어 있기 때문에, 교육시설이 많은 주택지역에서 주로 발생하는 유형이라 볼 수 있다. 군집 4는 도로시설 중 보행시설이나 문화교육 시설 중 교육시설, 서비스시설 중 금융시설이 인접한 곳에서 주로 발생하는 유형이라 볼 수 있다.

표 8은 보다 상세화된 공간적 진술에서 개념수준1의 공간 객체와의 연관성을 나타낸다. 군집 1은 연관된 규칙이 발견되지 않았고, 일반적인 공간적 진술에서와는 달리 군집 3, 4는 임계치를 만족하는 강한 연관성을 찾을 수 없었다. 도로시설과 교차하면서 산업시설이 근처에 있으면 군집 2이며, 근처에 산이 있는 것처럼 고도가 주위보다 높으면 군집 2 유형의 특성을 나타낸다는 규칙만 도출되었다. 이것은 군집 2와 공간 객체의 일반화된 공간적 관계와는 사뭇 다른 규칙을 보여준다.

표 9는 상세화된 공간적 진술에서 개념수준2의 공간 객체와의 연관성을 나타내고 있는데, 군집 1과 군집 2와 연관을 나타내는 규칙은 도출되지 않았다. 군집 3은 일반화된 공간적 진술에서 교육시설과 주택이 군집 3과 강하게 연관된 것과 같이 보다 세부적으로 주택이 근처에 있다면 군집 3의 특성을 나타낸다는 규칙이 도출되었다. 군집 4는 매우 개별적인 연관 규칙을 나타내고 있어 일반적인 위상적 계층과는 다른 연관 규칙들을 보여주고 있다.

5. 결론

서울시 강남구 교통사고 데이터에 공간 데이터 마이닝의 적용한 결과, 많은 결과물을 얻을 수 있었다. 하지만 이러한 결과들은 모두가 유의미하거나 흥미로운 규칙들일 수 없다. 따라서 본 분석의 결과는 맥락에 따라 다양하게 해석될 수 있으며, 보다 심화된 연구를 위한 새로운 가설들로 사용될 수 있을 것이다. 교통사고의 공간적 분포 패턴은 다음과 같다.

교통사고의 발생 위치는 강남구 전체에 걸쳐 도로 네트웍 상에 골고루 분포하고 있지만, 발생 지점마다 다른 발생 건수를 나타냈다. 특히 강남대로의 논현역, 양재역 부근, 역삼역 부근, 개포 근린공원 일대는 매우 빈번하게 교통사고가 발생하고 있다.

군집 분석을 통해서 4개의 군집 유형이 확인됐으며, 군집마다 다른 사고특성을 나타내고 있다.

위치상으로 강남대로의 논현역, 테헤란로의 역삼역 부근의 사고 발생 빈도가 가장 높은 지점들이 군집 1의 유형을 나타냈고, 두 번째로 빈도가 높은 압구정로와 삼성로가 교차하는 지점, 도곡역 부근, 역삼로 개나리 아파트 부근 역시 군집 1의 유형을 나타냈다. 한 지점에서 높은 빈도를 나타내지는 않지만, 한남대교 일대의 사고들도 군집 1의 유형이 대부분 차지한다. 전체적으로 강남구 교통사고의 가장 많은 유형을 반영하듯이 사고가 잦은 지점들이 주로 군집 1의 유형을 띤다고 볼 수 있다. 군집 1의 특성은 강남구 교통사고의 전체적인 경향성을 많이 반영하고 있다. 이것은 군집 1의 유형이 강남구 교통사고 전체의 44%로 가장 많은 유형이기 때문으로 추정된다. 한편 전체적인 경향성과 달리 나타나는 특징적인 특성은 사고일반에서 사고내용이 전체적인 경향과 달리 중상사고의 빈도가 가장 높은 특성을 가지고 있고, 물적 피해액이 전체 평균보다 높다. 사고특성에서는 신호등이 없을 때 발생한 특성을 가지고 있다. 연관 분석을 통한 결과는 교통사고 개별 속성 사이의 연관을 나타내는 규칙은 나타나지 않았으며, 공간 객체와의 공간적 연관성도 발견되지 않았다. 즉 군집 1의 유형은 개별 속성이나 공간 객체와의 연관을 보이지 않으며, 데이터에서 흥미로운 규칙이 발견되지 않는다.

군집 2는 압구정로 일대와 학동로, 남부순환로 일대에서의 주로 발생하는 사고 유형이며, 선릉로와 봉은사로 일대를 따라서는 매우 선형적으로 발생하는 사고 유형이다. 또한 논현역과 역삼역 부근, 삼성역과 압구정로와 삼성로 교차로 부근의 사고와 같은 빈도가 높은 사고도 군집 1의 유형과 더불어 많이 나타나는 유형이다. 군집의 특성은 군집 1의 유형과 유사하게 전체적인 경향성과 매우 흡사하며, 단지 사고일반에서 사고자 연령이 전체 평균보다 약간 낮은 계층에 의해서 발생했으며, 사고특성에서 군집 1처럼 신호등이 없는 곳에서 주로 발생한 사고들이다. 연관 분석을 통해 발견된 규칙은 개별 속성들에서 중앙분리 표시, 속도 규제와 같은 도로환경적 요인과 강한 연관을 가지는 유형이라는 것을 보여준다. 또한 이웃한 공간 객체들도 이 유형의 특성에 영향을 미치는데, 일반화된 공간적 진술에서는 개념수준1의 객체와 유의미한 규칙이 발견되지 않으며, 서비스와 문화교육, 의료후생과 같은 가장 추상화된 수준에서 연관성을 보인다. 보다 상세화된 공간적 진술에서 역시 개념수준1의 객체와는 연관성이 없으며, 개념수준2의 객체, 특히 자연적 객체(산)가 근처에 있으면 군집 2의 특성을 나타낸다.

군집 3은 전체 강남구 교통사고 발생에서 가장 적은 유형으로 한남대교 부근과 남부순환로 도곡역 일대에서 사고특성을 보여준다. 군집의 특성은 사고일반에서 사고 발생월이 전체적인 경향과 달리 5월과 같은 봄철에 주로 발생했던 사고 유형들이며, 물적 피해가 각 군집 중에서 가장 낮은 유형이다. 봄이라는 계절적 요인은 물적 피해액이 크거나 중상자가 상대적으로 많이 발생하는 겨울철 사고에 비해 물적 피해가 적은 경미한 사고의 특성을 반영한다고 추정될 수 있다. 연관 분석의 결과에서 볼 수 있듯이 속성들 중 주로 노면상태나 속도 규제와 같은 도로환경적 요인이 군집 3과 강한 연관을 보이고 있다. 이웃한 공간 객체도 군집의 유형과 강한 연관을 보이는데, 서비스, 의료후생, 주택, 문화교육, 도로시설이 일반화된 공간적 진술에서 연관을 보이고 있다. 개념수준2에서는 서비스 시설 중 교육시설과 주택이 함께 나타나는 곳에서의 발생 특성을 보이고 있다. 보다 상세화된 공간

적 진술에서는 개념수준2의 객체 주택이 근처에 있으면(nearby) 군집 2라는 결과를 보여주고 있다.

군집 4는 강남대로 양재역 부근, 개포 근린공원 부근, 양재대로 삼성역 부근의 높은 빈도의 사고 중 대부분을 차지하는 유형이고, 세곡동 세곡교 부근과 같은 강남구 외곽지역에서 발생한 유형이라 볼 수 있다. 군집 4는 사고일반에서 주로 다른 군집과 달리 상대적으로 금요일과 같은 주말에 많이 발생한 특성을 가진다. 한편 사고특성에서는 전체적인 경향과 가장 다른 형태의 특징들을 보여주는데, 통행목적에 있어서 개인용무 중의 사고였으며, 도로형태상 교차로에서 주로 발생했고, 차도폭은 3m 이하의 좁은 차로에서 발생했으며, 주위의 토지이용은 비시가지였다. 강남구 대부분의 간선도로가 평균 27m의 매우 넓은 대로이며, 부심으로서 대부분이 시가지임을 감안할 때, 간선도로보다 소로나 외곽지역으로 이동 시에 발생한 사고들로 추정할 수 있다. 또한 사고요일이 주말이며, 통행목적이 개인용무가 많은 걸로 봐서 주말 피크닉을 즐기려고 이동하는 사람들에 의한 사고 유형이라는 것을 추정할 수 있다. 연관 분석 결과는 개별 속성 중 도로환경적 요인과 다소 연관을 보여주며, 공간 객체와의 공간적 연관성에서는 개념수준1에서 문화교육 시설이 공통적으로 존재하면서, 동시에 서비스, 도로시설, 의료후생과 같은 인공물들이 함께 이웃해 있으면 높은 연관성을 나타낸다. 특히 개념수준2에서는 도로시설 중 보행시설이나 문화교육 시설 중 교육시설, 서비스시설 중 금융시설이 인접한 곳에서 주로 발생하는 유형이라 볼 수 있다. 하지만 위상적 계층이 보다 상세화된 수준에서는 뚜렷한 규칙을 찾아볼 수 없다.

공간 데이터마이닝을 이용한 교통사고의 공간적 패턴 분석은 몇 가지 한계점을 가지고 있다. 먼저 교통사고라는 현상이 교통 네트워크 상에서 발생한다는 특성과 지리적으로 연속적인 특성을 고려치 못하고 있다. 이것은 공간 데이터마이닝의 토대가 되는 공간 데이터베이스 구축에서 발생하는 한계라고 볼 수 있으며, 보다 정교하고, 정확한 교통사고 데이터베이스의 구축을 통해 이를 보완해야 할 것이다. 또한 공간 객체와의 연관성에 관한 연구에서 일반적인 공간 객체와의 연관성만을 다루고 있다. 향후의 연구는 교통사고와 강한 연관성을 가질 수 있는 공간 객체 설정에 보다 세심한 주의가 필요할 것이다.

註

1) 오재학, 이대근(1995)은 전산 입력되는 데이터는 조사자와 통계원표 작성자의 불일치, 89가지의 세부 코드로 된 통계원표 양식의 복잡성 등의 원인으로 부정확성을 가질 수 있다고 언급했다.

2) 일반적인 데이터마이닝의 작업에서 군집 분석(cluster analysis)은 전통적으로 통계학에서 많이 사용해오던 기법으로 데이터베이스의 객체들을 의미있는 하위 클래스(군집)로 그룹화시키는 것이다. 그룹화의 기준은 각 객체들이 가지는 비공간적 속성에 대한 다변량 통계치일 수도 있으며, X, Y 좌표를 변수로 하는 공간적 통계치일 수 있다. 전자와 같은 군집 분석은 엄밀한 의미에서 개념적 군집(conceptual clustering)[2]이라고 말할 수 있으며(Han and Fu, 1996, 409), 후자는 공간적 군집(spatial clustering)이라 볼 수 있다.

3) 공간적 분류(spatial classification)는 객체의 속성 뿐 아니라 이웃에 있는 객체들의 속성도 고려하여 분류하는 것을 말한다(Ester *et al*, 1997, 57). 분류를 위한 방법은 여러 가지가 있지만, 가장 일반적으로 사용되는 것은 의사결정 나무이다.

4) 데이터베이스의 데이터 셋이 가지는 종속성 분석을 위해서는 주어진 데이터 셋에서 함께 자주 나타나는 속성값의 조건을 보여주는 연관 분석(association analysis)이 주로 사용된다. 연관 분석은 흔히 장바구니 분석이라고 하며, 주어진 데이터 셋에서 속성들 사이의 흥미로운 관계를 찾는 분석이다. 연관 분석은 지지도(support)와 신뢰도(confidence)라는 기본개념을 가지는데, 도출되는 규칙들은 최소 지지도, 최소 신뢰도와 같은 사용자 정의 임계치를 통해서 결정된다. 공간적 연관 분석은 일반적인 연관 분석의 공간적 확장이며, 공간 객체들의 공간 관계를 통하여 연관 규칙을 찾는다. 공간 객체들의 공간 관계는 *adjacent_to, near_by, inside, intersect*와 같은 공간적 진술(predicate)로 표현된다. 공간적 진술은 위상 관계뿐 아니라, 방향, 거리 관계도 포함한다. 공간적 연관 분석은 이러한 공간적 진술을 포함하는 객체 관계를 표현하게 된다.

5) 공간적 이례 분석은 공간적 분포에서 전체적인 경향을 벗어난 이례 지점들을 탐색하는 것이다. 이러한 공간적 이례 분석은 시·공간적 경향에 따른 실시간 시스템 모니터링에서 사용이 가능하고, 민감한 지역에서의 테러리스트, 강도들의 움직임의 탐지에 매우 유용하다. 이례 분석에 대한 연구는 통계적 영역에서 대부분 수행되어 왔다.

6) 데이터의 서브 셋에 대한 압축 기술인 요약화(summarization)는 대상 데이터 객체의 특성을 비공간적 속성과 이웃한 객체들의 속성에 의해서 기술하는 공간적 특성화와 비공간적 속성과 공간적 속성을 특정한 개념 계층

에 의해서 요약하고, 보다 추상적인 수준에서 표현하는 공간적 일반화가 있다. 두 가지 모두 보다 넓은 의미에서 일반화를 기반으로 하는 공간 데이터마이닝으로 간주될 수 있다.

7) 경찰청의 『도로교통 안전백서』, 『교통사고 통계』, 도로교통안전관리공단의 『교통사고 통계분석』 등이 있다.

8) 동일한 좌표의 생성은 실제로 같은 위치에서 여러 번 반복하여 발생했을 수도 있지만, 통계원표 작성시 작성자의 오류와 같은 좌표 기입의 에러로 인한 것일 수 있다. 하지만 데이터의 신뢰성에 대한 검정은 본 연구에서 다루고자 하는 것이 아니기 때문에 포함하지 않는다.

9) CrimeStat1.0을 이용하여 발견했고, 이 CrimeStat1.0 은 K-Means 알고리즘을 이용하여 군집을 발견한다. 각 중심점으로부터 1 표준편차에 있는 사례들을 포함하도록 하였다. 군집의 형태는 타원형(ellipse)을 취한다.

10) 각 군집의 중심점의 좌표와 사고가 잦은 지점들과 거의 동일하다.

11) Koperski *et al.*(1995)가 제시한 것처럼 '두 객체가 이웃한다(close_to)'와 같은 공간적 진술(predicate)이 '두 객체가 인접한다(adjacent to)', '근처에 있다(nearby)', '교차한다(intersect)' 같은 세부적인 공간적 진술의 하위 개념을 포섭하고 있는 계층을 말한다.

12) Koperski *et al.*(1995)에 의하면 공간적 진술은 하나의 진술에서 여러 개의 진술이 있을 수 있다. 하나의 진술을 1-predicate라 하고 K개의 진술들의 결합을 K-predicate라 정의했다. K-predicate에서 효율적인 연관 분석 알고리즘은 apriori 알고리즘이 사용된다.

13) 일반화된 진술을 coarse predicate라 하고, 상세화된 진술을 refined predicate라 한다.

文 獻

고상선 · 오석기, 1996, "교통사고 특성과 발생지점의 유형화에 관한 연구," 동아대학교 대학원 논문집, 21, 481-500.

고상선, 1996, "교통사고 야기 영향요인간의 상관성 분석에 관한 연구," 국토계획, 31(4), 225-226.

김정현 · 이수범 · 박병정, 2002, 교통사고 잦은 지점 및 구간 선정방법 개선에 관한 연구, 교통개발연구원.

김효종 · 서채연, 1995, "교차 교통량 특성이 교통사고에 미치는 영향에 관한 연구-광주시를 대상으로-," 국토계획, 31(2), 255-266.

노맹석, 2000, "통계적 기법을 이용한 데이터마이닝," 기초과학, 4(1), 91-98.

손은정 · 강인수 · 김태완 · 이기준, 1998, "클러스터링 분석에 의한 공간데이터마이닝 방법," 한국정보과학회 가을 학술논문집, 25(2), 161-163.

오재학 · 이대근, 1995, "지리정보시스템을 이용한 교통사고 분석의 과학화 방안," 교통개발연구원.

이주형 · 손동혁 · 윤문교, 1990, "교통사고 발생특성과 그에 따른 사고요인 분석에 관한 연구," 국토계획, 25(1), 135-154.

장남식 · 홍성완 · 장재호, 1999, 데이터 마이닝, 대청미디어.

Chen, M. S., Han, J. and Yu, P. S., 1996, Data mining: overview from database perspective, *IEEE Transactions Knowledge and Data Enginee-ring*, 1-40.

Ester, M., Kriegel, H. P., Sander, J. and Xu, X., 1996, A density-based algorithms for discovering clusters in large spatial databases with Noise, *Proc. 2nd International conference on Knowledge Discovery and Data Mining (KDD '96)*, AAAI Press, 226-231.

Ester, M., Kriegel, H. P., Sander, J. and Xu, X., 1997, Density-connected sets and their application for trend detection in spatial databases, *Proc. 3rd International Conference on Knowledge Discovery and Data Mining (KDD ' 97)*, AAAI Press, 10-15.

Ester, M., Kriegel, H. P., and Sander, J., 1997, Spatial data mining: a database approach, *Proc. 5th International Symposium on Large Spatial Databases(SSD ' 97)*, Springer, 47-68.

Ester, M., Frommelt, A., Kriegel, H. P. and Sander, J., 1998, Algorithms for characterization and trend detection in spatial databases, *Proc. 4th International conference on Knowledge Discovery and Data Mining (KDD '98)*, New York City, 44-50.

Ester, M., Kriegel, H. P. and Sander, J., 2001, Algorithms and applications for spatial data mining, In Miller, H. J., Han, J.*(eds), Geographic Data Mining and Knowledge Discovery*, Taylor and Francis, 160-187.

Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P., 1996, From data mining to knowledge discovery: an overview, in Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Ulthurusamy. R.*(eds.), Advanced In Knowledge Discovery and Data Mining, Cambridge*, MA: MIT Press, 1-34.

Gahegan, M., 2001, Data mining and knowledge discovery in the geographical domain, National Academies White Paper, *Intersection of Geospatial Information and Information Technology*, 1-8.

Han, J. and Fu, Y., 1996, Attribute-oriented induction in data mining, in Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Ulthurusamy, R.*(eds), Advanced In Knowledge Discovery and Data Mining, Cambridge*, MA: MIT Press, 399-421.

_____, 1999, Discovery of multiple-level association rules from large databases, *IEEE Transactions on Knowledge and Data Engineering*, 11(5), 1-8.

Han, J. and Kamber, M., 2001, *Data Mining: Concept and Techniques*, Morgan Kaufmann.

Kaufman, L., and Rousseeuw, P., J., 1990, *Finding Groups in Data: an Introduction to Cluster Analysis*, John Wiley & Sons.

Koperski, K., and Han, J., 1995, Discovery of spatial association rules in geographic information databases, *Proc. 4th Symposium on Large Spatial Databases(SSD '95)*, Portland, 47-66.

Koperski, K., Han, J., and Stefanovic, N., 1998, An efficient two-step method for classification of spatial data, *Proc. International Symposium on Spatial Data Handling (SDH ' 98)*, Vancou-ver, Canada, 45-54.

Matheus, C. J. and Chan, P. K., Piatetsky-Shapiro, G., 1993, Systems for knowledge discovery in database, *IEEE Transactions Knowledge and Data Engineering*, 5, 903-913.

Miller, H. J. and Han, J., 2000, Discovering geographic knowledge in data rich environ-ments: a report on a specialist meeting, *ACM SIGKDD Explorations*, 1(2), 105-107.

Whitelegg, J., 1987, A Geography of road traffic accidents, *Transactions of the Institute of British Geographer*, New Series, 12(2), 161-176.

최초투고일 04. 04. 30
최종접수일 04. 06. 18