대한지리학회지 제33권 2호 1998 (325~338)

선형 점자료에 있어서의 시·공 복합 군집의 탐색

홍 상 기*

Detecting Space-Time Clusters in Linear Point Data

Sang-Ki Hong*

요약 : 본 연구에서는 시·공 복합적인 선형 점 자료를 대상으로 시간과 공간을 함께 고려했을 때 자료 내에 군집(cluster)—시·공 복합 군집(space-time cluster)—이 존재하는 가를 검증하는 방법에 대해 논의하고, 실제 교통사고지점의 분포자료를 분석하여 군집의 유무를 통계적으로 검증하였다.

통계 분석의 결과 다음과 같은 사실이 확인되었다. 첫째, Knox의 분할표 방법과 Mantel의 역수 변환을 이용한 일반화된 회귀분석방법 모두 임계 거리 및 임계 시간 간격의 선택이 분석결과에 영향을 미친다. 둘째, 이러한 임의성을 극복하기 위해 다양한 임계 거리 및 임계 시간 간격(혹은 부가 상수)에 대해 반복 실험한 결과, 일부 임계값의 조합에서 시간과 공간이 서로 독립적이라는 귀무가설을 기각할 수 있는 증거가 발견되었다. 세째, 시·공 복합 군집의 파악에 가장 적합한 임계 거리와 임계 시간 간격은 공간적으로는 7000m, 시간적으로는 14일 혹은 21일이다. 마지막으로, 통계 분석과정에서 자료에 존재하는 중복 기록 사고들의 존재가 밝혀짐으로써 시·공 복합 군집 검증이 탐험적 자료 분석(exploratory data analysis)의 도구로서 기지는 가치를 확인할 수 있었다.

주요어 : 시·공 복합 군집, 선형 점 자료, 점 패턴 분석, 교통사고 자료

Abstract : This study discusses the methods for statistical test of the presence of space-time clusters in a complex linear-point data. It also provides the test procedures and the test results for alcohol-related traffic accident data.

The findings from the statistical analyses are as follows. First, in both the Knox's and the Mantel's method, it was verified that the result of the analyses are influenced by the choice of critical values. Second, repeated tests for space-time clustering found that, for certain combinations of critical values, there is enough evidence to reject the null hypothesis of no space-time clustering. Third, the critical interpoint distance that produced the evidence of space-time clustering was identified as 7000m, while the critical time intervals were 14 days and 21 days. Finally, a hidden fact in the data (the double counted cases) was revealed during the first phase of statistical analysis, which proves the value of the statistical test of space-time cluster as an exploratory analytical tool.

Key Words : spatio-temporal cluster, linear point data, point pattern analysis, traffic accident data

1. 서 론

점의 패턴, 즉 어떤 사건들의 위치가 점으로 지도화된 것은 우리가 접하는 가장 흔한 공간패턴 중의 하나이다. 축척에 관계없이 이러한 점 패턴들의 예는 흔히 발견된다. 숲속의 특정 식생의 분포, 새 둥지의 위치, 금속 또는 암석에 있는 결함의 위치, 은하계의 혹성들, 도시들의 분포, 지진발생지점 등(Ripley, 1981)은 모두 점으로 표현되어 공간상의 점 패턴을 형성할 수 있다. 따라서 지리학을 비롯하여 생태학, 전염병학, 생물학, 천문학, 지질학 등 여러 학문 분야에서 점 패턴에 대한 연구가 주요 연구 분야 중의 하나로 자리잡고 있다.

점 패턴 지도는 두 가지 요소로 구성된다. 하나는 점들(규명 하고자 하는 사건 또는 객체들)이고 또 하나는 그 점들이 차지하는 지리적 영역이다 (Boots and Getis, 1988). 자료의 성격에 따라 연구 대상 지역은 1차, 2차 혹은 3차원 공간이 될 수 있

* 국토개발연구원 책임연구원(Associate Research Fellow, Korea Research Institute for Human Settlements)

다. 예를 들어 고속도로상의 진출입지점들의 분포를 연구하는 경우 연구 지역은 1차원의 선이 될 것이다. 반면에 어떤 지역에 있어서의 정주입지에 대한 분석은 2차원 연구지역을 필요로 한다. 때에 따라서는 지진연구에서와 같이 3차원 연구지역을 설정하는 것이 보다 적절한 경우도 있다.

점 패턴에 대한 연구는 통계학, 특히 공간통계학(spatial statistics)의 분야에서 활발히 이루어져 왔다. 그러나 대부분의 점 패턴에 대한 연구는 2차원의 연구지역에 분포하는 점의 패턴을 분석하는 데 집중되어 왔다. 소아기 암(childhood cancer) 발생의 공간패턴에 대한 분석이 그런 예이다. 반면에 점들의 분포가 하나 또는 여러 개의 선에만 국한되는 경우 그 분포에 관한 분석은 거의 이루어지지 않은 것이 사실이다. Roder(1974)는 이렇게 점 현상의 발생가능 지역이 선상으로 국한되어 있는 점들에 대해 '선형 점'(linear point pattern)이라는 용어를 사용하였다. 이러한 선형 점들의 예로는 도로상 교통사고지점의 분포, 송전네트워크에서 발생한 고장지점의 위치, 강을 따라 여러 군데 지점에서 측정한 오염 수준의 분포, 주요 도로를 따라 위치한 상업활동의 분포 등을 들 수 있다.

선형 점 패턴에 있어서도 패턴분석의 복잡도는 자료의 성격과 연구목적에 따라 달라진다. 점 자료가 점들의 속성에 대한 고려 없이 단순히 위치만을 표시하는 것이라면 점 패턴은 단순히 기하학적으로 분석될 수 있으나, 만일 자료가 공간적, 그리고 비공간적(aspatial) 속성을 모두 포함하고 있다면 분석은 훨씬 더 복잡해 질 수 밖에 없다. 또한 자료가 공간 뿐 아니라 시간요소까지 포함하고 있을 경우 문제는 더욱 복잡해 진다. 왜냐하면 이 경우 점 패턴에 대한 분석이 단순히 공간영역에서만이 아니라 시간과 공간이 함께 고려되어 이루어져야 하기 때문이다.

이렇게 매우 복잡한 시·공 복합적인 선형 점 자료의 좋은 예가 교통사고지점의 분포이다. 이 연구에서는 이러한 다차원적인 선형 점 자료를 대상으로 시간과 공간을 함께 고려했을 때 자료 내에 군집(cluster)—시·공 복합 군집(space-time cluster)—이 존재하는 가를 검증하는 방법에 대해 논의하고 실제 교통사고지점의 분포자료를 분석하여 군집의 유무를 통계적으로 검증하였다.

2. 점 패턴에 있어서의 시·공 상호작용의 분석

공간상의 점 패턴은 공간자료 중 가장 단순한 형태의 것이다. 이 패턴에 관심을 갖는 것은 관찰된 점의 분포가 어떤 체계적인 경향을 보이는 가를 알기 위함이다. 이러한 점 패턴의 분석을 위해 공간통계학에서는 많은 분석 방법들을 개발하였다. 그 중에는 공간을 구역별로 나누고 그 속에 포함된 점의 개수를 세는 단순한 방법(quadrat method)부터 공간상의 점 프로세스를 모델링(spatial point process modeling)하는 것까지 여러 가지 방법이 있다[1].

그러나 여기서 주목할 점은 공간통계학적 기법의 대부분이 2차원의 연구지역을 대상으로 사용되어 왔을 뿐 선형 점 패턴에 적용될 수 있는 통계적 이론이나 기법은 찾아보기 어렵다는 점이다. 이러한 제약을 극복하기 위하여 이 연구에서는 2차원 공간을 대상으로 한 연구에서 주로 적용되는 두 점 사이의 직선거리(euclidian distance) 대신 네트워크상에서 두 점을 연결하는 최단경로거리(shortest path distance)를 사용하였다. 앞에서 언급된바와 같이 선형 점들은 선 위에서만 발생될 수 있으며, 이는 네트워크의 일부분이다. 네트워크의 형태와 사건들의 발생위치에 따라서 직선 거리는 두 점들간의 네트워크상 실제 거리를 심각하게 왜곡시킬 수 있다. 따라서 사건들 간의 최단경로거리를 분석에 사용함으로써 자료에 존재하는 시·공 복합 군집의 증거를 확인하기 위한 통계적 분석이 더욱 현실적인 것이 될 수 있다.

공간통계적 기법을 확장하여 시·공 상호작용(Knox, 1964)의 효과를 통계적으로 검증하는 데는 여러 방법이 있다. 그 중 한 가지 방법은 '완전한 공간적 무작위성'(complete spatial randomness, CSR)의 개념을 '완전한 공간-시간적 무작위성'(complete spatio-temporal randomness, CSTR: Cressie, 1993)의 개념으로 확장하는 것이다. 이때 CSTR은 공간뿐만 아니라 시간차원에서도 자료내에 어떠한 구조도 포함되어 있지 않음을 의미한다. 이 개념은 따라서 시·공 상호작용을 검증하는 데 필요한 귀무가설을 제공하게 된다.

시·공 상호작용의 분석에 자주 사용되는 또 하

나의 방법은 자료내에 존재하는 시·공 복합 군집의 유무를 확인하는 방법이다. 만일 시간적으로 인접한 사건들 중 공간적으로도 가까운 사건들이 존재하고, 그 확률이 우연에 의해서 발생할 확률보다 크다면, 시·공 복합 군집이 존재한다고 할 수 있다(McAuliffe and Afifi, 1984).

시·공 복합 군집의 정도를 측정하는 방법으로는 여러 가지가 있다. 예를 들어 Knox(1964)는 모든 가능한 점들의 쌍을 시간 차원과 공간 차원에서 동시에 고려하여 시·공 복합 군집의 유·무를 검증하였다. Knox의 방법은 여러 가지 질병 발병 지점의 공간적 패턴에 시·공 복합 군집이 존재하는 지를 연구하는 데 많이 사용되었으며, 그 중 가장 대표적인 것이 Burkitt의 임파육아종(Burkitt's Lymphoma)의 시·공 복합 군집성에 관한 연구(Doll, 1978)이다. 또한 Mantel(1967)은 점의 분포에 관한 Knox의 가정(Poisson 분포)을 제거하여 일반화된 회귀분석방법(generalized regression approach)을 제시하였으며, 분석에 사용되는 거리의 척도로서 단순히 사건들 사이의 시간, 공간적 거리를 사용하는 대신에 점들 간의 시간적, 공간적 거리를 좀 더 충실하게 반영할 수 있는 근접성 척도(closeness measure)로서 거리의 역수(1/거리)를 사용할 것을 제안 하였다. 이밖에도 McAuliffe and Afifi(1984)는 근접성 척도로서 최소이웃거리(nearest neighbor distance)를 사용할 것을 주장하기도 하였다.

3. 자료 및 연구 대상 지역

이 연구의 자료로는 미국 North Carolana 주에 있는 Pitt County를 대상으로 1987에서 1989까지 수집된 교통사고 발생지점에 관한 자료를 사용하였다. Pitt County는 County의 가운데에 위치한 중규모 도시인 Greenville을 중심으로 몇 개의 소도시가 있는 전형적인 농촌지역으로, 대도시를 포함한 지역에 비해 교통사고 발생 건수가 상대적으로 적어 분석이 용이하다는 점과 여러 해에 걸쳐 매우 자세한 교통사고 지점 및 사고와 관련된 여러 가지 속성들이 조사되어 있다는 점에서 연구 지역으로 선정되었다. 그림 1은 North Carolina 주에서 Pitt County의 위치를 보여주며 그림 2는 Pitt County에 있는 주요 도시들의 위치와 주요 도로망을 보여준다.

연구지역에서 3년간 수집된 교통사고와 그 발생지점의 수는 모두 3,680건 이다. 이 모든 사건들을 공간적으로 표현했을 경우 교통사고의 성격상 과도한 공간적 중복이 일어나 시·공 상호작용에 대한 통계적 검증을 수행하기에는 부적합하다고 생각된다. 따라서 시·공 상호작용에 대한 분석은 비교적 사고 건수가 적고 시간, 공간적으로 중복이 덜한 음주와 관련된 교통사고(253건)에 대해 행하였다.

다음의 그림 3에서 그림 5까지는 음주운전과 관련된 교통사고의 분포를 '선형 점 자료의 시·공

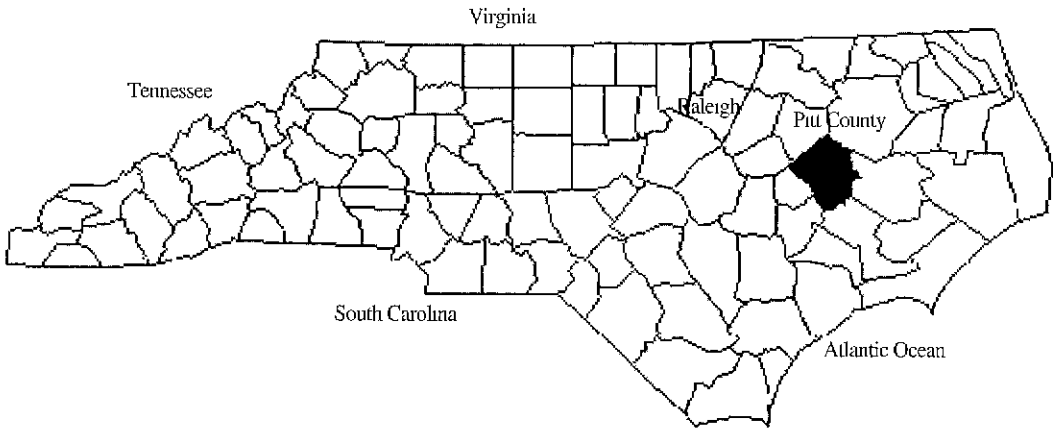

그림 1. 미국 North Carolina 주에서의 Pitt County의 위치

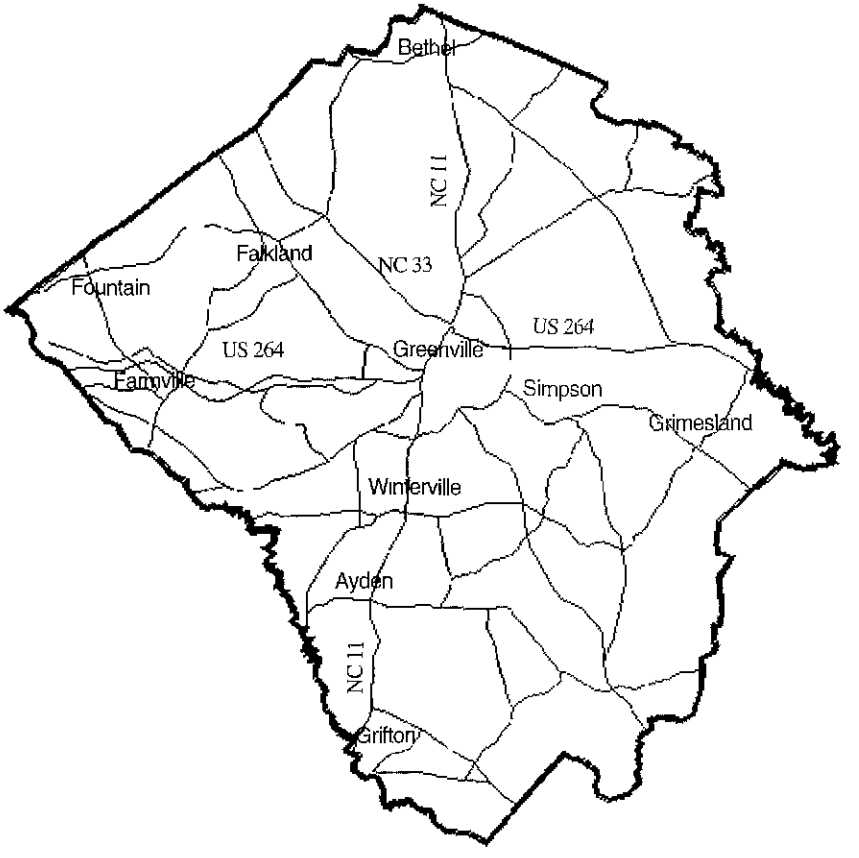

그림 2. Pitt County내의 주요 도시들과 주요 도로

간적 분석을 위한 대화형 시각화 시스템'(Hong, 1997)을 사용하여 나타낸 것이다. 이 시각화 시스템은 복잡한 시·공간 속성을 가진 자료에 존재하는 여러가지 숨겨진 패턴들을 시각적 방법으로 분석하기 위해 개발되었으며, 다양한 시각화 기법을 통해 점들의 분포와 속성을 나타낼 수 있어 교통사고 자료와 같이 복잡한 점 자료를 표현하는 데 적합하다.

그림 3은 Pitt County 전체를 대상으로 음주운전 관련 교통사고의 공간적 분포를 나타낸 것이다. 이 그림에서 각 점의 색깔은 교통사고의 정도, 다시 말해서 치사사고(적색), 부상사고(녹색), 그리고 대물사고(황색) 등을 표시하고 있다. 여기서 주목할 점은 county 수준에서의 분포를 나타내기 위해 사고지점을 표시하는 심볼(8면체)의 크기를 확대하였기 때문에, 사고지점들이 공간적으로 많이 중복되어 나타나고 있다는 점이다. 따라서 자세한 사고지점의 공간적 분포는 네트워크의 일부를 확대하고 점을 표시하는 8면체의 크기를 줄임으로써 살펴볼 수 있다.

그림 4는 이렇게 확대된 네트워크의 일부(Greenville 주변)에 사고지점들을 표시한 것이다. 또한, 교통사고의 시간적 분포는 3차원 공간의 수직축을 사용함으로써 시각적으로 분석 가능하다. 그림 5는 같은 지역을 대상으로 교통사고의 발생시점을 요일별로 분류하여 일요일부터 월요일까지 7개의 층으로 나타내어 그것을 3차원 공간에서 표현한 것이다. 이때 각 점의 색깔은 사고발생 요일을 나타내고 있어 요일별 구분을 시각적으로 할 수 있도록 도와준다. 위에서 언급한 시각화 시스템

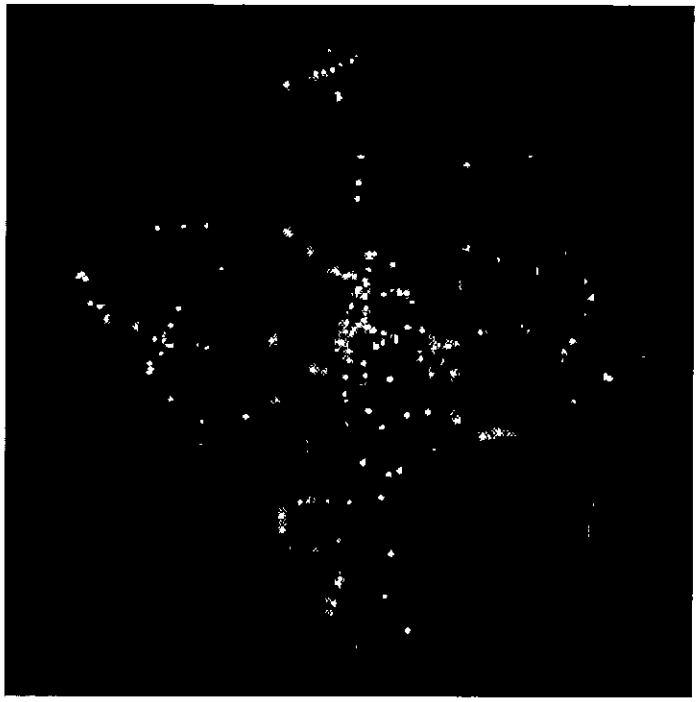

그림 3 Pitt County의 음주운전관련 교통사고의 공간적 분포

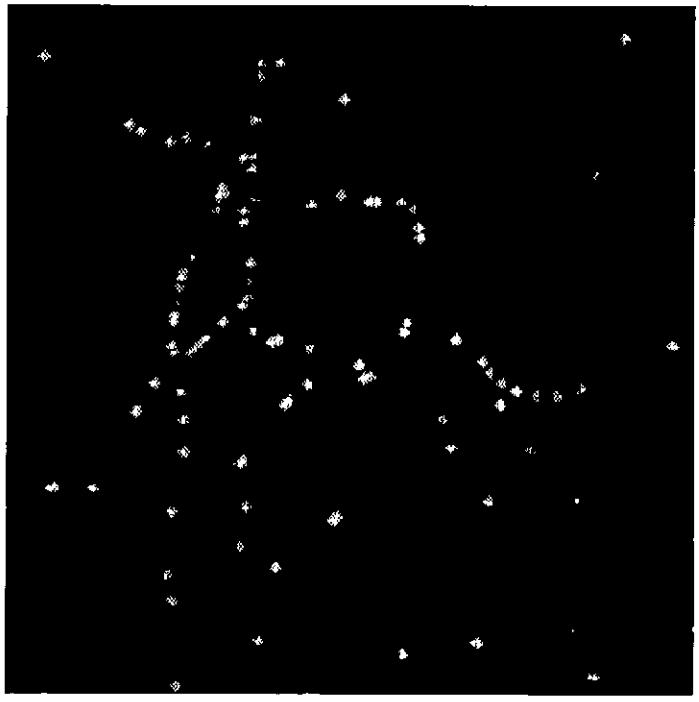

그림 4. Greenville 주변의 음주운전관련 교통사고의 공간적 분포

그림 5. 발생 요일별로 분류한 음주운전관련 교통사고의 분포

(Hong, 1997)을 사용하여 음주운전 관련 교통사고의 시·공간적 분포를 시각적으로 분석한 결과 일부 지역에서 시·공 복합적으로 군집이 나타나는 것을 발견할 수 있었다. 따라서 다음의 단계는 과연 자료내에 존재하는 이러한 시·공 복합 군집이 통계적으로 유의한가를 검증하는 것이다.

4. 시·공 복합 군집의 유무에 대한 통계적 검증

이 절에서는 시·공 복합 군집의 유무를 통계적으로 검증하기 위해 사용된 방법들에 대해 설명하고 이 방법들을 자료에 적용시켜 본 결과에 대해 기술한다. 통계적 검증에 사용된 방법은 Knox의 분할표 방법(contingency table approach)과 Mantel의 일반화된 회귀분석 방법(generalized regression approach)이다. 이 절에서는 또한 위의 방법들이 갖고 있는 제약조건들을 보완하기 위해 수행된 통계적 절차에 관해서도 설명한다.

1) Knox의 분할표 방법

Knox의 방법은 n개의 사건들을 관측한 값으로부터 형성된 n(n-1) 점들의 쌍을 고려한 값에 근거를 두고 있다. 시·공 복합 군집의 유무를 확인하기 위해서는 우선 시간적으로 가까우면서 공간적으로도 가까운 점들의 쌍을 찾을 필요가 있다. 이 경우에 '근접도'(closeness)는 이분법적이다. 다시 말해서, 두 사건이 미리 정해진 일정한 거리(또는 시간 간격)이내에 들기만 하면 이 두 사건은 서로 근접한 것으로 분류된다. 이러한 사건들의 쌍들은 2×2 분할표의 네 개 중 한셀에 위치된다. 만약에 공간적으로 근접한 X 쌍과 시간적으로 근접한 Y 쌍들이 있고, 시간과 공간의 독립성을 가정할 수 있다면, 시·공간적으로 동시에 근접한 쌍의 갯수(Z)는 평균 $XY/n(n-1)$을 가지는 포와송(Poisson) 분포를 보이게 된다. 만약 관측된 Z값이 이러한 분포에서 나온 기대값에 비해 유난히 크게 나올 경우 시간과 공간이 서로 독립적이라는 귀무가설을 버리고 시·공 복합 군집이 존재한다고 말할 수 있다.

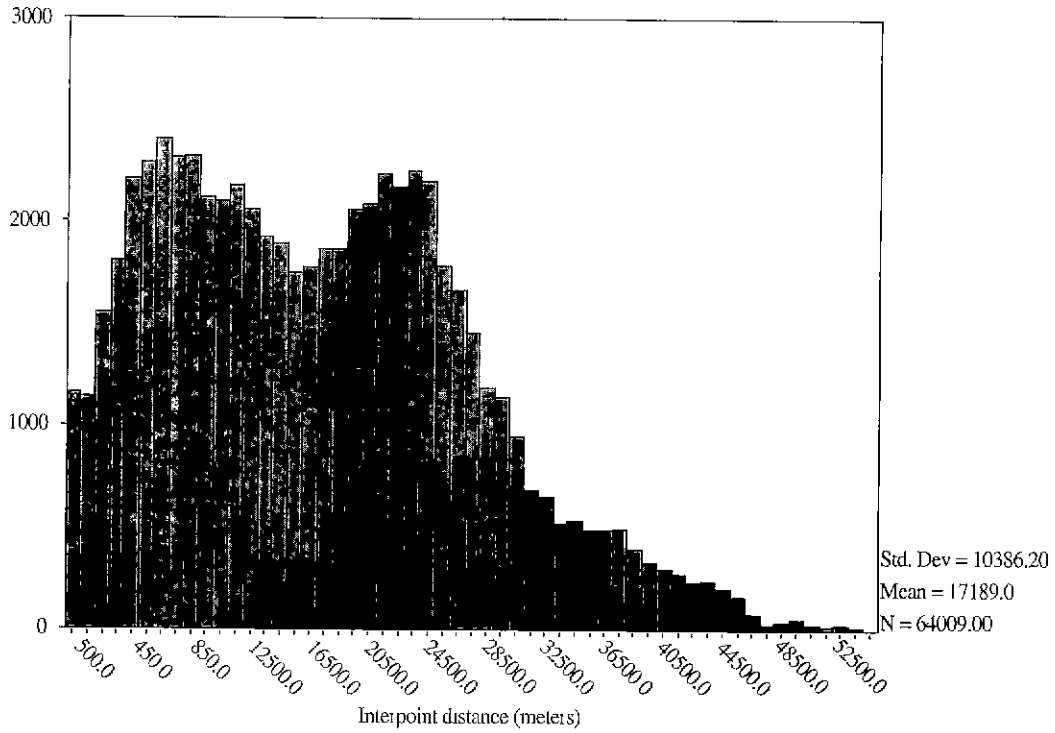

그림 6 사건쌍들 사이의 공간적 거리의 빈도 분포

Knox 방법의 문제점은 분석에서 임의의 임계 거리(critical distance)를 사용하는데 있다. 사용된 거리값의 선택에 따라 같은 자료로부터 상당히 다른 결과를 얻을 수도 있기 때문이다. 임계 거리의 선정에 대해 Knox는 자료의 성격을 면밀히 검토하여 분석하고자 하는 사건들에 대해 의미가 있는 값들을 이용할 것을 제안하였다(Knox 1964). 그러나 음주운전관련 교통사고의 경우에서와 같이 분석하고자 하는 대상에 대해 의미있는 임계 거리를 정하기가 모호한 경우가 있다. 이러한 경우에는 자료의 빈도분포를 살펴봄으로써 임계 거리에 대한 단서를 찾아 볼 수 있다. 또한 임계 거리를 달리하여 반복적인 분석을 수행함으로써 임의의 임계 거리를 사용한데서 발생되는 분석의 약점을 보완하는 것이 가능하다.

그림 6과 그림 7은 공간과 시간 각각에 대한 사건들 간의 거리의 빈도분포를 나타낸다. 그림 6에서는 공간적으로 거리 2000m와 7000m지점에 단절점(break point)이 존재함을 알 수 있다. 7000m 이내에 위치하는 두 사건이 공간적으로 '가깝다'라고 분류할 것인가는 연구대상지역의 지역적 범위를 고려하여 결정되어야 한다. 연구의 대상지역 Pitt county의 지리적 범위를 고려할 때, 공간 거리 7000m 이내는 공간적으로 가깝다고 할 수 있다는 것이 연구자의 판단이다.

거리 자료와는 달리 시간 간격의 빈도분포(그림 7)는 어떤 명백한 단절점을 제시하지 않는다. 그러므로 다양한 시간 간격을 사용하여 그 중 어떤 시간 간격에서 흥미있는 패턴이 나타나는 가를 확인하는 것이 합리적일 것으로 생각된다. 공간차원에서와 마찬가지로 사건들이 시간적으로 가깝다, 멀다를 결정하는 시간 간격을 임계 시간 간격(critical temporal distance)이라 부르기로 한다. 분석에 사용된 임계 시간 간격들은 0, 7, 30, 14, 21, 28, 30, 90, 365 등 이다. 이 중 0은 동일한 날에 사고가 발생된 경우를 나타내며, 7의 배수는 주 단위의 순환적 경향을 보기위해, 30은 월 단위를, 90은 계절을, 그리고 365는 년 단위를 나타낸다.

Knox의 분할표 방법을 자료에 적용한 결과를 나타낸 것이 표 1이다. 분석에 사용된 자료의 수는

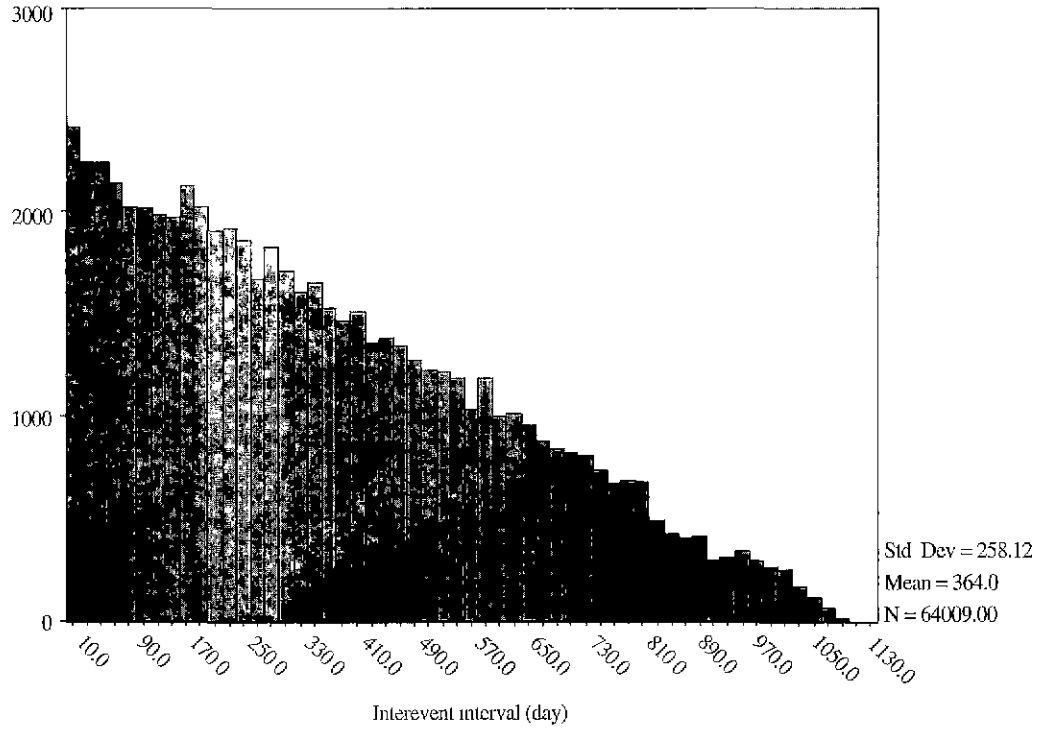

그림 7. 사건쌍들 사이의 시간적 간격의 빈도 분포

표 1 Knox의 분할표 방법을 이용한 분석 결과

Distance (m)		Interval(days) 0	7	14	21	28	30	90	365
500	Exp	0.6068	6.6438	12.899	18.92	25.253	27.027	76.894	276.86
	Obs	*16*	*24*	*28*	*36*	*46*	*46*	*96*	268
1000	Exp	1.1109	12.162	23.613	34.636	46.229	49.476	140.77	506.84
	Obs	*16*	*26*	*38*	*48*	*60*	*60*	*152*	468
1500	Exp	1.7275	18.913	36.72	53.861	71.889	76.938	218.9	788.17
	Obs	*16*	*30*	*42*	*64*	*80*	*80*	*220*	760
2000	Exp	2.5031	27.406	53.207	78.046	104.17	111.48	317.19	1142.1
	Obs	*18*	*36*	*62*	*92*	*112*	*114*	*324*	1108
7000	Exp	15.038	164.65	319.66	468.88	625.82	669.78	1905.6	6861.3
	Obs	*30*	*186*	*350*	*504*	*632*	*676*	*1918*	6860

주) 진한 이텔릭으로 표시된 숫자는 관측값이 기대값을 초과하는 경우를 나타냄

253건 이다. 표 1에서 알 수 있듯이 Knox의 방법을 사용한 분석 결과는 대부분의 시간 및 공간 임계 거리의 조합에서 실제 관측값이 기대값 보다 현저히 높음을 보여준다. 따라서 이들 임계 거리의 조합에서 시·공 복합 군집이 존재하지 않는다는 귀무가설을 잠정적으로 기각할 수 있다. 왜 잠정적인 기각인가에 대한 이유는 다음에 설명한다.

특히 임계 거리 500m가 사용될 경우 0부터 30까지의 모든 임계 시간 간격에 있어 관측값이 기대값을 크게 상회하고 있다. 이것은 시·공 복합 군

집의 존재에 대한 강한 증거라고 볼 수 있으며 이런 경향은 분석에 사용되는 공간 임계 거리가 증가할수록 약해지나, 임계 거리 2000m와 임계 시간 간격 7일에 이르기까지 나타나고 있음을 알 수 있다.

시간 측면에서 분석 결과를 살펴보면 관측값과 기대값 간의 가장 큰 격차는 임계 시간 간격이 0일(동일한 날의 사건)이 사용될 경우에 나타남을 알 수 있다. 이 임계 시간 간격에서는 임계 거리 7000m까지 모든 임계 거리에서 시·공 복합 군집의 강한 증거가 나타난다. 임계 시간 간격이 증가할수록 그 증거는 모든 임계 거리 영역에 있어 감소하나, 작은 임계 거리를 사용했을 때가 여전히 큰 임계 거리를 사용했을 때보다 더욱 강한 복합 군집의 증거를 보인다.

위의 결과가 비록 시·공 복합 군집이 존재한다는 강한 증거를 제시하고는 있지만, 자료내에 존재하는 패턴을 좀 더 자세히 이해하기 위해 기대값들과 관측값들간의 격차가 어떻게 변화하는 가를 살펴보았다. 가장 먼저 눈에 띄는 것으로는 기대값과 관측값의 격차가 현저히 크게 나타나는 경우가 동일한 날의 사고를 시간적으로 근접한 사건들로 생각했을 때(간격 0)라는 점이다. 따라서 왜 이런 패턴이 나타나는 가를 규명하기 위해서는 같은 날 발생된 사고들을 좀 더 자세히 검토할 필요가 있다.

동일한 날에 발생한 교통사고들로 시간 간격을 제한했을 경우 16개의 쌍들이 이 범주에 속한 것을 발견할 수 있었다. 이 16개 사고쌍들의 속성을 검사한 결과 사실은 8건의 음주관련 교통사고들이 각각 두번씩 기록된 것이라는 것을 알 수 있었다. 이 사고들이 두 번씩 기록된 이유는 사고에 개입된 두 운전자 모두 음주 운전을 하였으나 기록이 운전자를 중심으로 이루어져 별개의 사건으로 기록되었기 때문으로 보인다. 따라서 두 자료점들은 같은 위치에 동일한 시간에 나타날 수 밖에 없었던 것이다. 이것을 두개의 개별 사고로 간주할 지 아니면 단일사건으로 간주할 지는 연구자가 판단할 문제이다. 이번 연구에서는 이들 중복 기록된 사건이 시·공 복합 군집의 유무 검증에 미치는 영향이 매우 커서 결과적으로 전체적인 결과를 왜곡시킬 가능성이 많다는 점에서 단일 사건으로 처리하기로 하였다.

표 2는 중복 기록된 사건들을 단일사건으로 처리한 후 Knox의 방법에 의해 분석한 결과를 나타낸다. 이때 분석된 자료의 수는 245건 이다. 예상한 대로 여기서도 역시 몇몇 거리간격 조합에 대한 시·공 복합 군집의 증거가 나타나나 그 정도나 경향은 앞의 표 1의 경우보다 낮게 나타나고 있다. 500m를 임계 거리로 선택할 경우 임계 시간 간격 7, 28, 30과 90일의 경우에서 관측값이 기대값을 현저히 초과함을 알 수 있다. 2000m를 임계 거리로 사용하면 기대값보다 큰 관측값을 보이는 경우는 같은 날 발생한 사고들만을 시간적으로 근접한 사

표 2. 수정 자료를 대상으로 한 Knox의 분할표 방법을 이용한 분석 결과

Distance (m)		Interval(days) 0	7	14	21	28	30	90	365
500	Exp	0.4376	5.8639	11.407	16.658	22.274	23.806	67.712	243.76
	Obs	0	*6*	10	16	*26*	*26*	*72*	230
1000	Exp	0.827	11.082	21.558	31.482	42.096	44.991	127.97	460.68
	Obs	0	8	18	26	38	38	124	416
1500	Exp	1.2968	17.377	33.802	49.363	66.005	70.543	200.65	722.34
	Obs	0	12	22	42	58	58	192	690
2000	Exp	1.899	25.446	49.5	72.287	96.657	103.3	293.83	1057.8
	Obs	*2*	18	40	68	88	90	288	1012
7000	Exp	11.324	151.74	295.17	431.05	576.37	616	1752.1	6307.6
	Obs	*14*	*166*	*324*	*468*	*590*	*630*	*1758*	6280

주) 진한 이텔릭으로 표시된 숫자는 관측값이 기대값을 초과하는 경우를 나타냄

건으로 처리한 경우 한 가지 뿐이다. 7000m 임계 거리의 경우가 시·공 복합 군집의 증거가 가장 강하게 나타난다. 이 때는 임계 시간 간격 365일의 경우를 제외한 모든 시간대에서 관측값이 기대값을 초과하였다.

수정 자료를 가지고 수행한 Knox 분석의 결과와 원자료(중복 기록된 사건들을 두 개의 다른 사건으로 간주한 자료)로 부터의 분석 결과는 다음과 같은 차이를 보인다. 우선 시간적으로 동일한 날에 발생한 사건의 중요성이 현저하게 감소했다. 수정 자료들을 이용해서 수행한 시험에서는 동일한 날을 임계 시간 간격으로 사용했을 경우, 단지 두 개의 임계 거리(2000m와 7000m)에서만 기대값들을 초과하는 결과를 나타낸다. 원자료를 사용한 분석결과에서는 같은 날을 임계 시간 간격의 기준으로 삼았을 때, 임계 거리의 대, 소에 상관없이 모든 관측값이 기대값을 초과하였다.

둘째. 원자료의 분석결과보다 수정 자료의 분석결과에서 시·공 복합 군집의 증거가 약하다는 것이다. 게다가, 원자료의 결과는 많은 임계 거리 및 임계 시간 간격에서 시·공 복합 군집의 증거가 발견되어 특정한 거리를 임계 거리로 삼기에 어려움이 있으나, 수정 자료는 7000m라는 임계 거리가 시·공 복합 군집의 증거를 나타내는 데 가장 의미있는 거리라는 것을 확실히 보여주고 있다. 이에 관해 우리는 사건들간의 거리의 빈도 분포에서 7000m가 주요 단절점의 하나였다는 데 주목할 필요가 있다(그림 6 참조). Knox 방법을 이용한 분석 결과를 고려해 볼 때, 만약 서로 7000m이내에서 발생한 사건들을 공간적으로 근접한 사건으로 간주했을 경우, 자료내에는 시·공 복합 군집의 강한 증거가 있음을 나타낸다고 할 수 있다.

표 2에서 확인할 수 있는 다른 사실은 자료의 분포에 관한 Knox의 가정(포와송 분포)에 무리가 없다는 사실이다. 이론적인 기대값들과 관측값들이 매우 근접해 있다는 사실은 이를 잘 증명한다.

2) Mantel의 일반화된 회귀분석 접근

Mantel(1967)은 Knox의 테스트에서 사용되는 포아송 분포의 가정을 제거함으로써 Knox의 테스트를 확장시켰다. Knox의 방법과 같이 사건쌍들 간의 시간, 공간적 거리, 즉 관측값은 $Z=\sum\sum X_{ij}Y_{ij}$로 정의되며, 여기서 X_{ij}는 사건 i 와 j의 거리를 나타내며, Y_{ij}는 대응하는 시간 간격을 나타낸다. Z의 귀무 분포(null distribution)는 유한 모집단 접근법(finite population approach)을 통해 얻을 수 있다.

만약 공간상에 n개의 점이 있고 시간상의 n개의 점이 있다면 시·공 복합 군집이 없다는 가설(귀무가설)은 공간상의 점들이 시간상의 점들과 무작위로 대응한다는 것과 동일하며 따라서 그 결과 총 n!의 동일 확률을 가지는 쌍이 형성되게 된다. 그러므로 귀무 분포는 n!개의 가능한 순열(permutation)을 나열하고 각 순열에 대해 Z값을 계산하므로써 얻어지게 된다. Mantel(1967)은 표본 집단의 평균과 변량을 n!개의 가능한 모든 순열을 검토하지 않고도 계산할 수 있는 방법을 고안하였다[2]. 궁극적으로 관측값(Z), 기대값(Exp Z), 분산($Var\,Z$)값으로부터 Mantel의 통계치인 t 값을 구하면:

$$t = \frac{Z - ExpZ}{\sqrt{Var\,Z}}$$

로 표현된다.

표 3. Mantel의 방법을 이용한 Knox의 분할표 방법 검증

Distance(m)	Interval(days)							
	0	7	14	21	28	30	90	365
500	*14.0349*	*4.81197*	*3.02711*	*2.84587*	*3.0119*	*2.66728*	*1.68335*	-0.5146
1000	*10.066*	*2.8447*	*2.139*	*1.65174*	*1.48289*	1.09736	0.73404	-1.6316
1500	*7.76851*	*1.83497*	0.63253	1.00998	0.70387	0.25725	0.05787	-0.8941
2000	*7.04269*	1.18769	0.88016	1.16201	0.5681	0.17666	0.29945	-0.8503
7000	*3.04689*	*1.31721*	1.37353	1.33026	0.2035	0.19813	0.24486	-0.0085

주) 진한 이텔릭으로 표시된 숫자는 α=0.1(양방 검증의 경우 α=0.2)에서 유의함을 표시

만일 사건들이 시간적으로 그리고 동시에 공간적으로 근접해 있으면 X_{ij}와 Y_{ij}에 각각 1의 값을 주고, 그렇지 않을 경우에 0을 주게 되면, $Z=\sum\sum X_{ij}Y_{ij}$는 Knox의 분할표(contingency table)에서 시간과 공간 모두에서 근접한 쌍들의 수의 2배와 일치할 것이다. 따라서 X_{ij}와 Y_{ij}에 0 또는 1의 값만을 사용함으로써 Mantel의 통계를 사용하여 Knox의 분할표 방법이 자료 분석에 적합한 것인가를 검증할 수 있다. 표 3과 표 4는 원자료와 수정자료의 각각에 대해 Mantel의 방법을 사용하여 분석한 결과를 나타낸다.

Mantel의 방법을 사용한 분석 결과는 원자료(253개 경우)를 사용할 경우 Knox 분석의 결과와 유사하게 나타난다. 시·공 복합 군집의 강한 증거는 현저하게 짧은 공간적 거리와 짧은 시간 간격에서 발견된다. 군집에 대한 증거는 Knox 방법을 사용한 경우의 패턴과 매우 유사하다. Mantel의 방법에서 자료의 분포에 대해 어떠한 가정도 하고 있지 않는 데도 불구하고 통계치의 분포가 Knox의 방법을 사용했을 경우와 유사하다는 것은 Knox가 가정한 자료의 분포에 대한 가정(포와송 분포)이 합리적인 것이라는 것을 확인해 준다.

Mantel의 방법을 수정 자료(245건)에 대해 적용했을 경우, 그 결과(표 4)는 Knox 방법의 결과와 약간의 차이점을 보인다. Knox의 방법에서는 임계 거리 7000m에서 모든 임계 시간 간격에서 유의한 시·공 복합 군집의 증거가 발견되었으나 Mantel의 방법을 이용하면 임계 거리 7000m에서 단지 임계 시간 간격 7일과 14일이 사용되었을 때만 군집의 증거가 강하게 나타난다. 그러나 이 경우에도 역시 임계 거리 7000m에서 가장 시·공 복합 군집의 증거가 강하다는 점은 동일하다.

3) 역수 변환(Reciprocal Transformation)을 이용한 Mantel의 일반화된 회귀분석 방법

Knox의 분할표 방법의 장점은 어느 일정한 거리 또는 시간을 기준으로 그 안에 속하는 경우들을 동일한 값으로 처리하고 또 그것을 초과하는 모든 경우들을 같은 값으로 처리하여 공간적, 시간적 거리들을 다룰 때 발생하는 극단적인 수치의 변화를 피할 수 있다는 점이다. 그러나 Knox의 방법은 임계 거리 안에 들어 가는 사건들 간의 거리의 차이를 무시하고 있다는 점, 즉 사건들 사이의 실제 거리, 또는 시간 간격이라는 귀중한 정보가 유실된다는 단점이 있다. 그러므로 이러한 임계 거리의 사용을 피하기 위해, Mantel은 서로 아주 멀리 떨어진 사건들의 거리는 축소하는 반면 짧은 거리를 두고 발생한 사건들은 그 거리를 전개(spread out)할 수 있는 자료 변환을 제안하였다. 다음의 변환을 이용하면 0 과 1의 수치를 사용하는 대신 실제 거리 수치가 공간적 근접성을 측정하는데 사용될 수 있다. 변환식은 다음과 같다.

$$X_{ij}=\frac{1}{d_{ij}+k_s} \text{ 그리고 } Y_{ij}=\frac{1}{t_{ij}+k_t}$$

여기서 d_{ij}와 t_{ij}는 각각 i 사건과 j 사건 사이의 공간 거리, 시간 간격을 나타내며, k_s와 k_t는 동일한 공간 또는 시간 좌표를 갖는 사건들의 경우 분모가 0이 되는 것을 피하기 위한 임의의 상수이다.

역수 변환을 이용한 Mantel의 방법은 임계 거리를 사용하지 않아도 되는 반면, k_s와 k_t의 부가적인 상수의 선택이 분석 결과에 영향을 미친다. 왜냐하

표 4 Mantel의 방법을 이용한 Knox의 분할표 방법 검증(수정된 자료 사용)

Distance(m)	Interval(days)							
	0	7	14	21	28	30	90	365
500	-0.4605	0.0401	-0.2997	-0.1168	0.5758	0.3286	0.4026	-0.8667
1000	-0.6474	-0.6634	-0.5533	-0.7104	-0.4622	-0.7641	-0.2721	-1.993
1500	-0.8136	-0.9275	-1.4722	-0.7658	-0.7248	-1.1008	-0.4758	-1.084
2000	0.0526	-1.0667	-0.9847	-0.3707	-0.652	-0.9708	-0.2666	-1.2035
7000	0.6192	0.9095	*1.3473*	*1.4537*	0.4677	0.4651	0.1214	-0.1949

주) 진한 이텔릭으로 표시된 숫자는 $\alpha=0.1$(양방 검증의 경우 $\alpha=0.2$)에서 유의함을 표시

면 만약 선택된 상수가 매우 작다면, 거리가 0에 가까운 사건들의 영역이 과도하게 확장되어 군집이 존재하는 데도 불구하고 그것을 찾아내지 못할 수 있기 때문이다. Knox의 임계 거리를 설정하는 경우와 마찬가지로 역수 변환을 사용한 Mantel의 방법에 있어 부가적인 상수를 어떻게 선택해야 할 것인가에 대한 최적 기준은 알려진 바 없다 (McAuliffe and Afifi, 1984). 이에 대해 Mantel은 다양한 상수값을 적용하여 여러 번 분석을 시행하는 방법을 제안 하였다(Mantel, 1967).

표 5는 역수 변환을 이용한 Mantel의 방법을 다양한 부가 상수를 사용하여 적용한 분석 결과를 보여준다. 여기서는 원자료 대신 중복 기록된 사건이 제외된 수정 자료(245건)를 사용하였다. 표 5에서 k_s는 공간 거리에 대한 부가 상수, k_t는 시간 간격에 대한 부가 상수를 나타낸다.

수정 자료를 사용하여 유의도 수준 α = 0.05 (양방 검증의 경우 α= 0.1)에서 통계적 유의성을 가진 부가 상수의 세가지 조합을 발견할 수 있다. 다시 말해서 부가적인 거리 상수가 k_s = 0.1이고 부가적인 시간 상수가 k_t = 1, k_t = 14, k_t = 21의 경우 시·공 복합 군집의 존재에 대한 통계적으로 유의한 증거가 발견된다.

앞에서 행한 분석 결과에서 임계 거리 7000m와 임계 시간 간격 21일까지에서 시·공 복합 군집의 증거가 발견되었다는 점을 상기할 때, 역수 변환을 사용한 통계치의 분포가 Knox의 분할표 방법 그리고 Mantel의 회귀분석방법(0 과 1만을 이용한 경우)의 분석 결과와 유사한 패턴을 보이는 것은 흥미있는 사실이다.

앞에서도 설명한 바와 같이 작은 부가 상수를 사용하면 아주 가까운 거리에 있는 사건의 쌍의 존재가 통계치에 미치는 영향이 감소되는 효과가 있다. 그 이유는 역수 변환과정에서 분모가 작으면 결과적으로는 큰 수가 생성되기 때문이다. 이렇게 되면 분석에서 사용되는 거리의 영역이 확장되게 되어 공간적으로 가까운 사건들 사이의 거리를 벌리는 결과가 된다. 시·공 복합 군집의 증거가 오직 낮은 거리 부가 상수(k_s = 0.1)를 사용했을 때만 보인다는 사실은 작은 거리 부가 상수를 사용하여 공간적으로 아주 가까운 사건들의 영향력을 감소시켰을 때 시·공 복합 군집이 발견되었음을 의미한다. 다시 말해서 서로 7000m 이내에서 발생한 사건들을 공간적으로 가까운 사건으로 간주할 경우 음주 관련 교통사고 자료에는 시·공 복합 군집이 존재한다고 말 할 수 있다 이때 시·공 복합 군집을 찾아내는데 가장 적합한 임계 시간 간격은 14일과 21일이다.

5. 결 론

이 연구에서는 복잡한 시·공간 자료를 대상으로 그 속에 존재하는 시간과 공간의 상호작용에 의한 군집의 유무를 통계적으로 검증하는 방법에 대해 논의하였다. 특히 연구의 대상이 선상에서만 발생할 수 있는 선형 점자료라는 점에서 기존의 점 패턴 분석들과는 다른 분석방법을 적용하였다. 선형 점 자료는 선 위에서만 발생하므로 두 점간의 거리를 측정할 때 직선거리를 사용하면 네트워크의 형태에 따라서는 거리의 왜곡이 심하게 되어 결과적으로 비현실적인 통계 분석 결과가 나올 수 있다. 그러므로 이 연구에서는 공간적 근접성을 측정하는 데 있어 점들 간의 네트워크상의 거리를

표 5. 역수 변환을 이용한 Mantel의 일반화된 회귀분석 방법의 결과(수정된 자료 사용)

k_s	k_t				
	0.1	1	14	21	30
0.1	0.509754	***2.505961***	*2*	***1.800373***	1.6129
500	-0.14257	0.660443	-0.12194	-0.07972	-0.05071
1000	-0.03352	-0.20772	-0.16604	-0.12238	-0.08971
2000	0.0796	-0.12021	-0.09309	-0.05423	-0.02469

주) 진한 이탤릭으로 표시된 숫자는 α=0.05(양방 검증의 경우 α=0.1)에서 유의함을 표시

사용했다.

분석 대상인 253개의 자료 점들에 대해 가능한 모든 쌍을 계산하면 63,256개의 쌍이 되며 이 모든 경우에 있어서 두 점들 사이의 거리를 네트워크상의 최단 경로를 따라 구하는 것은 상당한 계산이 요구되는 작업이다. 고성능의 컴퓨터와 GIS를 활용하지 않았다면 이 작업을 수행하기란 사실상 불가능 했을 것이다. 이렇게 얻어진 점들 간의 실제 거리를 사용함으로써 그것에 기반한 통계분석의 결과가 보다 의미있는 것이 되었다.

음주운전관련 교통사고 자료에 대한 시·공 복합 군집의 유무에 대한 통계 분석결과 다음과 같은 사실이 확인 되었다. 첫째, Knox의 분할표 방법과 Mantel의 역수 변환을 이용한 일반화된 회귀분석방법 모두 임계값의 선택이 분석 결과에 영향을 미친다는 것이 입증되었다. Knox의 방법에서는 기준이 되는 시간 간격과 공간 거리, 그리고 Mantel의 방법에서는 시간, 공간 부가 상수의 선택에 따라 시·공 복합 군집의 유무에 대한 통계치가 달라진다.

둘째, 이러한 임의성을 극복하기 위해 다양한 임계 거리 및 임계 시간 간격(혹은 부가 상수)에 대해 반복 실험한 결과, 일부 임계값의 조합에서 시간과 공간이 서로 독립적이라는 귀무가설을 기각할 수 있는 증거가 발견되었다.

세째, 서로 다른 통계적 분석기법을 사용했음에도 불구하고 모든 결과에서 유사한 임계 거리, 임계 시간 간격의 조합이 통계적으로 유의미하게 나타남으로서 어떤 임계 거리와 임계 시간 간격이 자료내의 복합 군집의 파악에 가장 적합한 가를 찾아낼 수 있었다. 다시 말해서 공간적으로는 7000m, 시간적으로는 14일 혹은 21일을 각각 공간적, 시간적 기준으로 삼았을 때 자료내의 시·공 복합 군집의 증거가 가장 강하게 나타난다.

마지막으로 주목할 것은 통계 분석 과정에서 자료에 존재하는 알려지지 않은 사실이 드러나게 된 것이다. 연구자는 통계 분석을 수행하기 전까지는 양쪽의 운전자가 모두 음주상태에 있을 때 발생한 사고가 별개의 사고로 기록된다는 사실을 알지 못했다. 1차 통계 분석의 결과는 가까운 거리에서 발생한 사건들이 통계치에 미치는 영향이 예상을 훨씬 뛰어넘고 있음을 보임으로써 이런 경우에 대해 보다 세밀한 검토를 하게 만들었다. 그 결과 앞에서 언급한 중복 기록 사고들의 존재가 밝혀지게 된 것이다.

사실 다양한 속성을 가진 복잡한 시·공간 자료는 그 복잡성이 분석의 가장 큰 걸림돌이 되어 자료의 성격을 파악하기가 어렵고 그 속에 존재하는 숨겨진 패턴이나 사실들을 발견하기가 힘들다. 이러한 자료들에 대해 통계학에서는 다양한 각도에서 그 성격을 파악하고 자료 속에 내재하는 흥미있는 패턴을 발견하려는 의도에서 '탐험적 자료분석' (exploratory data analysis, Tukey, 1977)이 많이 이루어 진다. 본 연구에서 자료내부에 깊이 감추어져 있던 중복 기록 사고의 경우들이 통계 분석의 결과로 발견되었다는 사실은 이러한 탐험적 자료 분석 도구로서의 시·공 복합 군집 검정의 가치를 다시 한 번 확인해 주는 것이라고 할 수 있다.

註

1) 공간 통계학적 점 패턴 분석에 관해서는 다음을 참조: Ripley, 1981; Cressie, 1993; Bailey and Gatrell, 1995

2) 이에 관한 자세한 절차는 Mantel(1967)의 부록 1과 부록 2 참조

文 獻

Bailey, T.C. and Gatrell, A.C., 1995, *Interactive Spatial Data Analysis*, John Wiley & Sons, New York.

Boots, B.N. and Getis, A., 1988, *Point Pattern Analysis*, Sage Publications, Newbury Park, CA.

Cressie, N., 1993, *Statistics for Spatial Data*, John Wiley & Sons, New York.

Doll, R., 1978, An epidemiological perspective of the biology of cancer, *Cancer Research*, 38, 3573-3583.

Hong, Sang-Ki, 1997, *Development and Proof-of-Concept of an Interactive Visualization System*

for the Spatio-temporal Analysis of Linear Point Data, Ph. D. Dissertation, Department of Geography, The Ohio State University, Columbus, Ohio.

Knox, 1964, The Detection of space-time interactions, *Applied Statistics*, 13, 25-29.

Mantel, N., 1967, The detection of disease clustering and a generalized regression approach, *Cancer Research*, 27, 209-220.

McAuliffe, T. L., and Affifi, A. A., 1984, Comparison of a nearest neighbor and other approaches to the detection of space-time clustering, *Computational Statistics & Data Analysis* 2, 125-142.

Ripley, B.D., 1981, *Spatial Statistics*, John Wiley & Sons, New York.

Roder, W., 1974, Application of a procedure for statistical assessment of points on a line, *The Professional Geographer*, 26(3), 283-290.

Tukey, J.W., 1977, *Exploratory Data Analysis*, Addison-Wesley, Reading, MA.