이학석사 학위논문

시공간 모형화를 통한
지하철역 승차인원 예측

Predicting Subway Passenger Flows
By Spatio-temporal Modeling

2017년 08월

서울대학교 대학원

통계학과

김 민 우

Contents

List of Tables

List of Figures

Chapter 1

서론 (Introduction)

1.1. 공간 데이터 (Spatial data)

오늘 날 여러가지 통계적 방법론에 사용되는 많은 데이터 중 공간 데이터란 위치 정보를 포함한 자료를 말한다. 여기서 위치 정보는 위도-경도로 대표되는 자료 각각의 공간적 위치를 나타내는 변수이며 공간 데이터의 예로 위치에 따른 강수량, 와이파이 강도, 멸종위기 동물의 출현 횟수 등을 생각할 수 있다. 공간 데이터를 분석함에 있어 전통적인 방법들을 적용하기 보다는 위치 정보에 따른 의존성을 고려한 새로운 방법을 적용해 볼 수 있으며 이러한 관점에 따라 공간통계학은 통계학의 한 분야로 자리잡으며 발전해 왔다.

공간 데이터는 크게 다음과 같이 세 가지 종류로 나뉠 수 있다.

1. 연속형 데이터 (Continuous data, Geostatistical data)
2. 이산형 데이터 (Discrete data, Areal data)
3. 점과정 데이터 (Point process data)

여기서 연속형 데이터는 주어진 데이터가 포함하고 있는 위치 정보가 연속적인 데이터를 말한다. 본 논문에서는 서울 지하철역의 승차인원 자료에 연속형 위치 정보를 부여하고 시간에 따른 의존성도 고려한 시공간(Spatio-temporal) 통계 방법론을 적용해 보았다.

1.2. 시공간확률과정 (Spatio-temporal Stohastic Process)

시공간 데이터를 아래와 같은 확률과정의 실현값으로 생각할 수 있다.

$$Y(\boldsymbol{s},t), \quad (\boldsymbol{s},t) \in D \times T \subseteq \mathbb{R}^d \times \mathbb{R}, \quad d \geq 1 \tag{1.1}$$

이 때 T는 시간을 나타내고 D는 관심이 되는 연속 공간이다. 예를 들어 D는 서울시 전체가 될 수 있으며 이 때 $\boldsymbol{s}$는 서울시 안의 한 지점이 되는 것이다. 확률과정 $\{Y(\boldsymbol{s},t),\ (\boldsymbol{s},t) \in D \times T\}$의 평균 함수와 공분산 함수는 다음과 같이 정의된다.

$$\mu(\boldsymbol{s},t) = E[Y(\boldsymbol{s},t)]$$
$$Cov\big(Y(\boldsymbol{s_1},t_1), Y(\boldsymbol{s_2},t_2)\big)$$

시공간 데이터의 분석에서는 이러한 공분산 함수에 위치 정보와 시점에 따른 의존성을 표현하는 것이 핵심이 된다.

1.3. 연구 목표

서울에서는 현재 영등포구 여의도동 샛강역을 출발해 대방역, 여의대방로, 보라매역, 보라매 공원, 신림역을 거쳐 서울대 앞을 연결하는 신림선 경전철

공사가 2021년 개통을 목표로 진행중이다. 신림선과 같이 새로운 지하철역이 생긴다면 과연 시간에 따른 승차인원은 얼마나 될까? 라는 물음에서 본 연구가 시작되었다.

서울의 많은 지하철역 중 강남, 고속터미널, 홍대입구, 서울역, 신림 등에서 상대적으로 많은 사람들이 승차한다. 그리고 출퇴근 시간에 더 많은 사람들이 지하철을 이용한다. 각각의 지하철역에 위치 정보를 부여하고 이에 따른 의존성과 시간에 따른 의존성을 고려한 시공간적 통계 방법론을 적용하여 분석을 진행하고 새로운 지하철역의 승차인원이 얼마나 될 것인가에 대한 예측을 수행해 볼 것이다.

Chapter 2

데이터 설명 (data description)

공공데이터포털 (www.data.go.kr) 에서 제공받을 수 있는 최초의 데이터는 서울시의 각 지하철역 (또는 버스정류장) 에서 시간대별 승차, 하차 인원을 월단위로 기록한 데이터이다.

	A	B	C	D	E	F	G	H
1	사용월	호선명	지하철역	04시-05시	04시-05시	05시-06시	05시-06시	06시-07시
2	201511	2호선	종합운동장	15	0	2612	1217	9438
3	201511	2호선	삼성	154	0	5748	6186	10836
4	201511	2호선	선릉	181	3	7264	6963	15111
5	201511	2호선	역삼	94	3	5124	8301	11213
6	201511	2호선	강남	326	6	15693	14092	27120
7	201511	2호선	교대	15	1	3387	6382	15241
8	201511	2호선	서초	8	1	1731	4771	6799
9	201511	2호선	방배	17	2	3897	3124	12160
10	201511	2호선	사당	65	2	13080	3680	25176
11	201511	2호선	낙성대	69	2	12630	2257	28907
12	201511	2호선	서울대입구	5034	27	32403	2746	49729

Figure 2.1: 공공데이터포털에서 받은 최초의 데이터 형태

2.1. 데이터 전처리(data pre-processing)

공공데이터포털에서 받은 데이터에 원하는 분석을 수행하기 위해서는 몇 단계의 전처리 과정이 필요했다. 먼저 시공간 모형화(Spatio-temporal modelling)를 위해 2015년 1월부터 2017년 1월까지 147개 지하철역의 승차인원 데이터를 추출하고 각 지하철역에 위치 정보를 부여하기 위해 구글API를 활용하여 지하철역의 위도-경도를 생성한 뒤 이를 중부원점 좌표계로 변환하여 x-y 변수로 나타내었다. 또한 환승역인지 여부를 나타내는 변수를 추가하여 환승역의 경우 1로 아닌 경우 0으로 표시하였고 계절을 나타내는 변수도 추가하여 12월 ~ 2월은 겨울(1), 3월 ~ 5월은 봄(2), 6월 ~ 8월은 여름(3), 마지막으로 9월 ~ 11월은 가을(4)로 표시하엿다. 마지막으로 각 지하철역에서 가장 가까운 버스정류장 5개의 시간대별 승차인원 정보도 추출하였다. 4장에서 다루게 될 모형화에서 지하철역의 시간대별 승차인원은 Y가 되고 환승역 여부, 계절, 버스 승차인원은 공변량(Covariate) X의 역할을 하게 된다.

	A	B	C	D	E	F	G	H	I	J	K
1	date	station	x	y	in06	in07	in08	in09	in10	in11	in12
2	201501	강남	202442.5	444276.1	31080	60589	98091	80352	81682	114348	143766
3	201501	개포동	205839.3	443297.7	3543	10259	14681	8937	6387	6358	7345
4	201501	경마공원	200699	438287.5	763	1639	2918	2346	2180	2789	3946
5	201501	과천	199681.7	437090.8	6428	21744	22540	14929	11321	10800	11332
6	201501	관악	191898.3	435582.3	12731	36265	34784	17672	12880	12409	14749
7	201501	구로디지털	191285.9	442871.6	57934	156695	207995	130817	82821	76320	79908
8	201501	구반포	198817.1	444672	1567	4258	5838	4213	3976	4588	5588
9	201501	구산	192697.8	456863	9486	28761	34424	18764	13839	12388	14141
10	201501	국회의사당	192738.8	447629.1	3065	3704	6209	7436	8593	11657	14922
11	201501	금호	201395.3	449860.5	8571	24348	37344	24247	15875	14599	13832
12	201501	길음	202200	455956.7	35804	110873	116472	69023	46350	41719	43987
13	201501	낙성대	196764.2	441965	27430	95887	143927	80285	45841	42347	46590
14	201501	남영	197452.2	449010.3	5937	14224	19614	12863	11968	13241	16782
15	201501	남태령	199035.4	440535	626	1677	3339	2567	1961	1689	1633

Figure 2.2: 전처리된 데이터

2.2. 데이터 시각화(data visualizing)

2017년 1월 147개의 지하철역 중 가장 승차인원이 많은 지하철역은 강남이며 이외에도 고속터미널, 홍대입구, 서울역, 신림, 사당, 구로디지털단지, 선릉에서 승차인원이 큰 값을 나타내었다. 반면 남태령, 구룡, 서빙고 등에서 가장 적은 수의 승객이 승차하였다. 아래 지도에서 점들은 각 지하철역의 위치를 나타내며 승차인원이 많은 지하철역일수록 크고 진한 점으로 표시되었다.

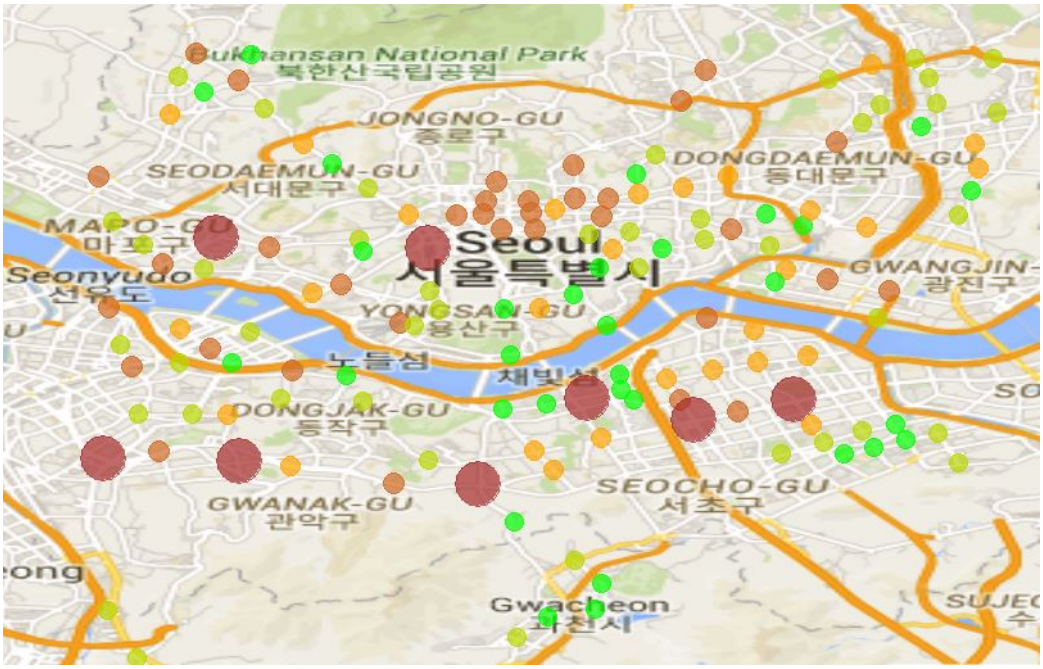

Figure 2.3: 지하철역의 위치와 상대적인 승차인원을 보여주는 지도

그렇다면 사람들이 지하철을 가장 많이 이용하는 시간은 언제일까? 6시부터 24시까지 1시간 단위로 측정하였을 때 147개 지하철 역 중 강남, 구로디지털단지, 여의나루를 비롯한 67개의 역에서는 18시에 가장 많은 승차인원을 기록하였고 신림, 낙성대, 미아사거리 등 63개의 지하철역에서는 오전 8시에

가장 많은 사람들이 지하철을 이용하였다. 반면 유일하게 제기동역에서만 15시에 가장 많은 사람들이 승차하였다. 이는 사람들이 출근시간과 퇴근시간에 지하철을 많이 이용함을 보여준다.

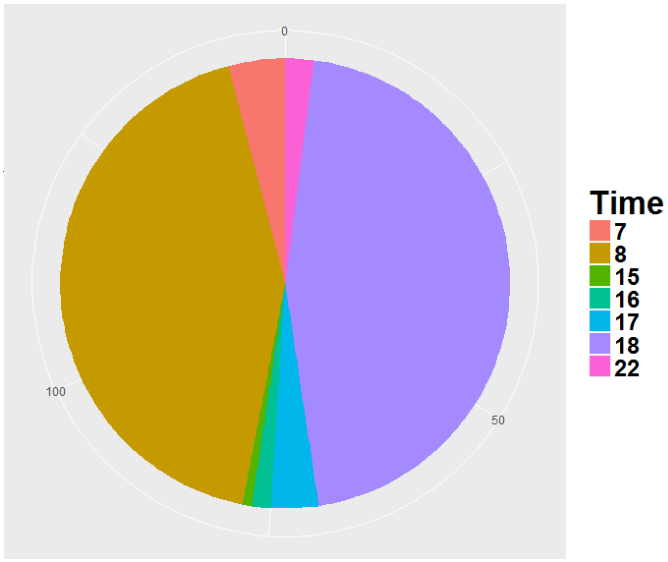

Figure 2.4: 147개 지하철역에서 가장 많은 수의 승객을 기록하는 시간이 언제인지 보여주는 파이 차트

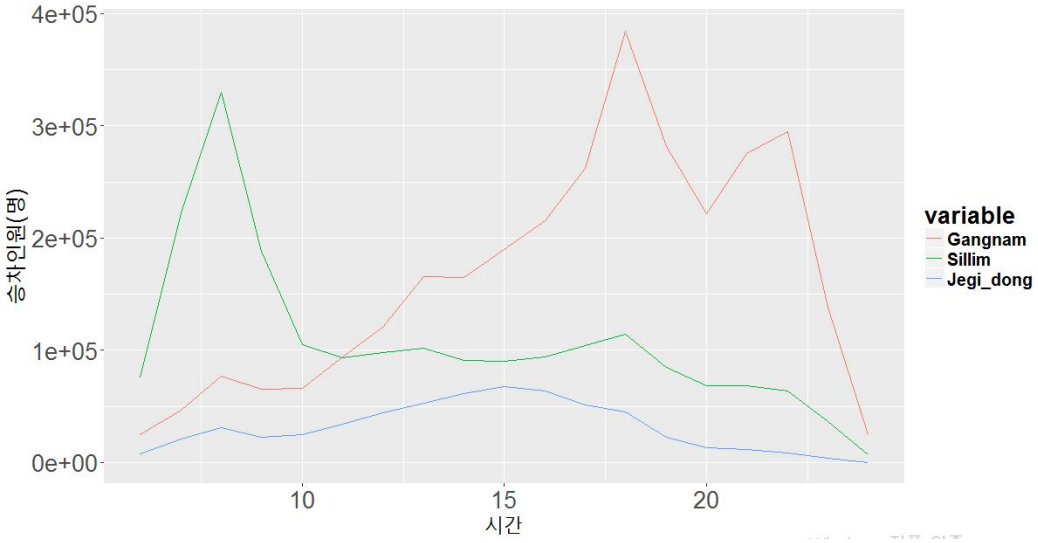

Figure 2.5: 강남, 신림, 제기동역의 시간에 따른 승차인원

다음으로 승차인원에 대한 히스토그램(Histogram)을 보면 오른쪽으로 꼬리가 긴 형태를 보인다. 따라서 로그변환(Log transformation)을 하여 분석을 진행할 필요가 있다.

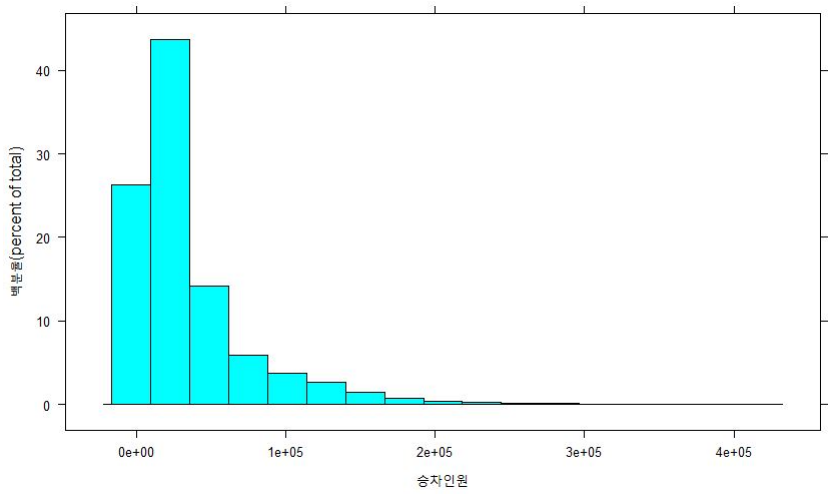

Figure 2.6: 승차인원 히스토그램

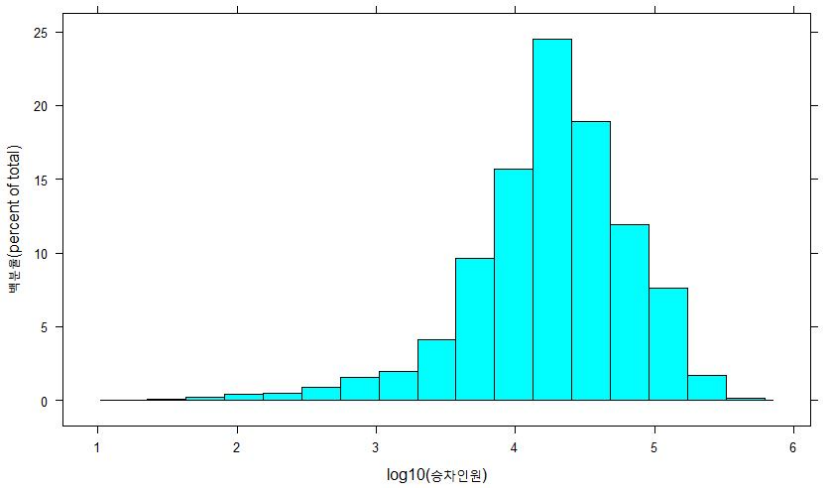

Figure 2.7: 로그변환된 승차인원 히스토그램

Chapter 3

의존성의 탐색 (Exploring dependencies)

데이터를 분석함에 있어 배리오그램 (Variogram) 은 우리가 가진 자료들의 의존성을 탐색하는 유용한 도구이다. 이 장에서는 이론적 베리오그램 (Theoritical variogram) 과 확률과정 $\{Y(\boldsymbol{s}),\ (\boldsymbol{s}) \in D\}$ 의 실현값 (Realization) 을 바탕으로 한 배리오그램 구름 (Variogram cloud) 와 경험적 배리오그램 (Emperical variogram) 을 설명하고 우리가 가진 지하철 데이터의 배리오그램을 구해보도록 한다.

3.1. 배리오그램 (Variogram)

배리오그램 (Variogram) 이 잘 정의되기 위한 조건인 고유 정상성 (Intrinsic stationarity) 조건은 아래와 같다.

$$(i)\ E[Y(\boldsymbol{s}+\boldsymbol{h})-Y(\boldsymbol{s})]=0, \quad \forall \boldsymbol{s},\ \boldsymbol{s}+\boldsymbol{h}\in D$$

$$(ii)\ Var(Y(\boldsymbol{s}+\boldsymbol{h})-Y(\boldsymbol{s}))<\infty, \quad \forall \boldsymbol{s},\ \boldsymbol{s}+\boldsymbol{h}\in D$$

고유 정상성은 증량 (Increments) 의 분산이 유한한 값을 가지며 두 지점 사이의 차이인 $\boldsymbol{h}$ 에만 의존함을 뜻한다. 고유 정상성 조건 하에서 배리오그램은

$$2\gamma(\boldsymbol{h}) := Var(Y(\boldsymbol{s}+\boldsymbol{h})-Y(\boldsymbol{s})) = E\big[(Y(\boldsymbol{s}+\boldsymbol{h})-Y(\boldsymbol{s}))^2\big]$$

로 정의되며 $\gamma(\boldsymbol{h})$ 는 세미배리오그램 (semi-variogram) 이라고 한다. 고유 정상성 보다 더 강력한 조건인 이차 정상성 (Second-order stationarity) 은 $Y(\boldsymbol{s})$ 가 상수 기댓값을 가지며 $Y(\boldsymbol{s})$ 와 $Y(\boldsymbol{s}+\boldsymbol{h})$ 의 공분산이 오직 $\boldsymbol{h}$ 에만 의존함을 말한다. 즉,

$(i)\ E[Y(\boldsymbol{s})]=\mu, \quad \forall \boldsymbol{s}\in D$

$(ii)\ C(\boldsymbol{h}) := Cov(Y(\boldsymbol{s}), Y(\boldsymbol{s}+\boldsymbol{h})) = Cov(Y(\boldsymbol{0}), Y(\boldsymbol{0}+\boldsymbol{h})), \quad \forall \boldsymbol{s},\ \boldsymbol{s}+\boldsymbol{h}\in D$

이며, 이 때 배리오그램과 공분산 함수 $C(\boldsymbol{h})$ 사이에는 다음과 같은 식이 만족한다.

$$\gamma(\boldsymbol{h}) := C(\boldsymbol{0}) - C(\boldsymbol{h})$$

시공간 확률과정 $\{Y(\boldsymbol{s},t),\ (\boldsymbol{s},t)\in D\times\mathbb{R}\}$ 의 이차 정상성 조건도 비슷하게 다음과 같다.

$$C(\boldsymbol{h_s}, h_t) := Cov(Y(\boldsymbol{s},t), Y(\boldsymbol{s}+\boldsymbol{h_s}, t+h_t)), \quad \forall(\boldsymbol{h_s}, h_t),\in D\times T$$

3.2. 배리오그램 구름 (Variogram cloud)

우리가 가진 n 개의 관측치, 즉 $\big(y(\boldsymbol{s_1}), y(\boldsymbol{s_2}), \ldots, y(\boldsymbol{s_n})\big)$ 로부터 배리오그램을 추정하는 것이 다음 단계이다. 첫 번째로 다음과 같은 두 지점 $\boldsymbol{s_i}$ 와 $\boldsymbol{s_j}$ 의 비유사도 (Dissimilarity) 를 생각해보자.

$$\gamma^*(\boldsymbol{h}) := \frac{\big(y(\boldsymbol{s_i}) - y(\boldsymbol{s_j})\big)^2}{2}$$

모든 가능한 서로 다른 $(\boldsymbol{s_i}, \boldsymbol{s_j})$ 쌍 $\frac{n(n-1)}{2}$ 개에 대하여 비유사도 $\gamma^*(\boldsymbol{h})$ 를 구하여 $\|\boldsymbol{h}\|$ 에 따라 점으로 찍어 표현한 것을 배리오그램 구름 (Variogram cloud) 이라고 한다. 8시와 18시의 승차인원에 대한 배리오그램 구름을 그려보면 다음과 같다.

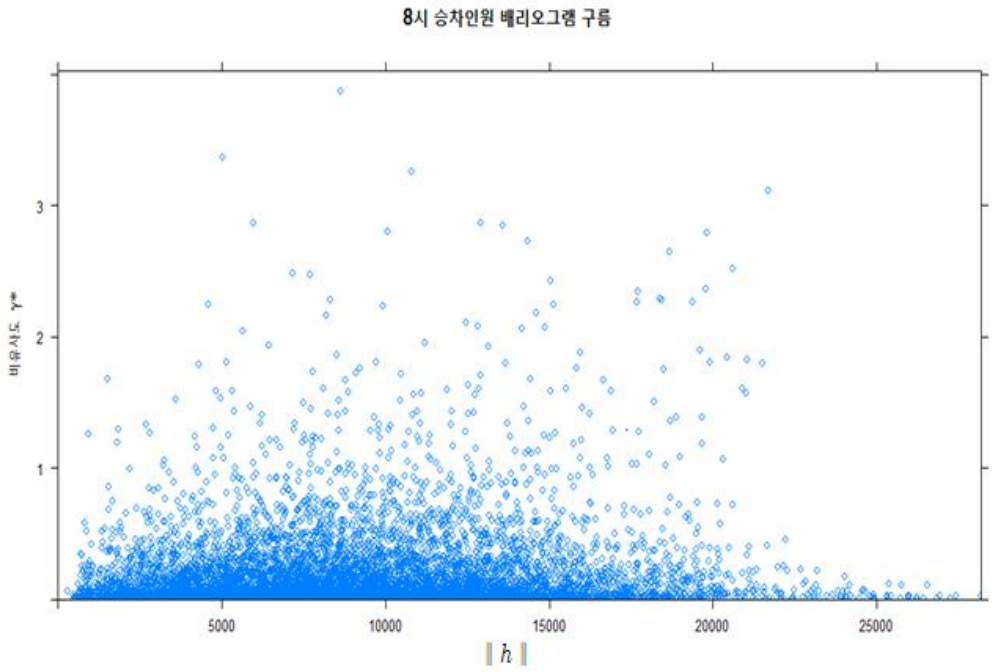

Figure 3.1: 8시 승차인원 배리오그램 구름

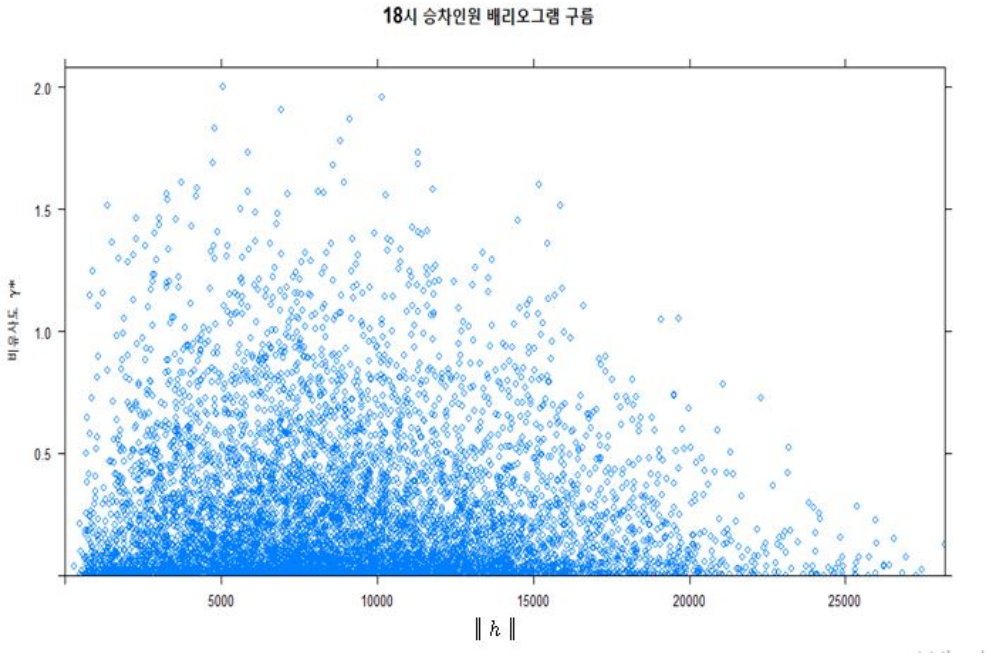

Figure 3.2: 18시 승차인원 배리오그램 구름

3.3. 경험적 배리오그램 (Empirical variogram)

배리오그램 구름을 통해 우리가 가진 데이터의 공간적 의존성에 대한 첫 통찰을 얻을 수 있지만 많은 점들이 밀집되어 표현되므로 배리오그램 함수의 실제 구조에 대한 유의미한 정보를 얻어내기는 어렵다. 따라서 거리 $\boldsymbol{h}$에 따라 몇 개의 대표값들로 표현되는 경험적 배리오그램 (Empirical variogram) 대해 알아볼 필요가 있다. 먼저 고전적 추정량 (Classical estimator) 또는 적률추정량 (Method-of-moment estimator) 은 다음과 같이 정의 된다.

$$\gamma^*(\boldsymbol{h}) := \frac{1}{2|N(\boldsymbol{h})|} \sum_{N(\boldsymbol{h})} (y(\boldsymbol{s_i}) - y(\boldsymbol{s_j}))^2$$

여기서 $N(\boldsymbol{h})$는 차이가 $\boldsymbol{h}$가 되는 모든 $(\boldsymbol{s_i}, \boldsymbol{s_j})$의 쌍들의 집합이다. 즉 $N(\boldsymbol{h}) := \{(\boldsymbol{s_i}, \boldsymbol{s_j}) : \boldsymbol{s_i} - \boldsymbol{s_j} = \boldsymbol{h} \quad for \quad i, j = 1, 2, \cdots, n\}$ 이며 $|N(\boldsymbol{h})|$는 $N(\boldsymbol{h})$의

원소의 수이다. 그런데 이 경우에는 $|N(\boldsymbol{h})|$가 0이 될 수도 있다. 따라서 가능한 모든 $\boldsymbol{h}$를 포함하는 K개의 분할 $\{H_k,\ k = 1, \ldots, K\}$를 생각하자. $M = \max_{i,j} \|\boldsymbol{s_i} - \boldsymbol{s_j}\|$라고 하면 $H_k = \{\boldsymbol{h} : (k-1)M/K < \|\boldsymbol{h}\| \leq kM/K\}$가 된다. $N(H_k) = \{(\boldsymbol{s_i}, \boldsymbol{s_j}) : \boldsymbol{s_i} - \boldsymbol{s_j} \in H_k \quad for \quad i, j = 1, \ldots, n\}$ 라고 할 때 다음과 같이 경험적 배리오그램을 정의할 수 있다.

$$\gamma^*(H_k) := \frac{1}{2|N(H_k)|} \sum_{N(H_k)} (y(\boldsymbol{s_i}) - y(\boldsymbol{s_j}))^2, \quad k \in \mathbb{N}$$

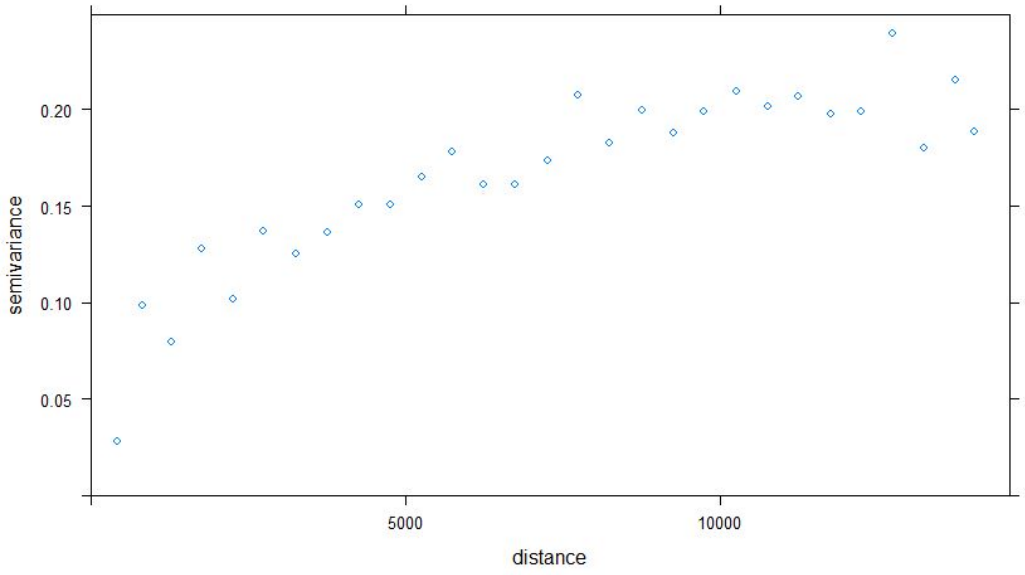

Figure 3.3: 8시 승차인원 경험적 배리오그램

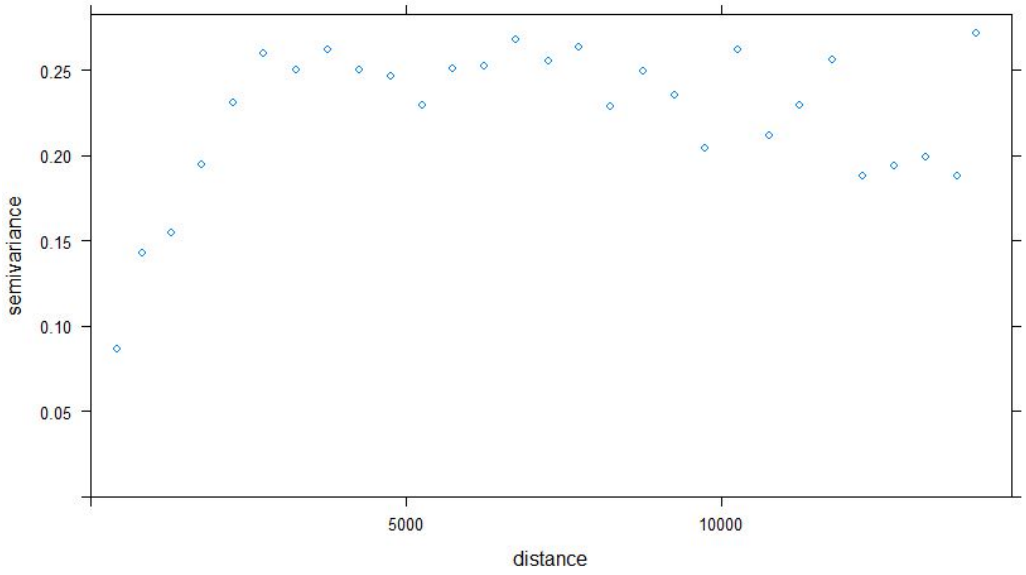

Figure 3.4: 18시 승차인원 경험적 배리오그램

시간과 공간에 따른 의존성을 함께 고려한 시공간 배리오그램(Spatio-temporal variogram)도 그려볼 수 있다.

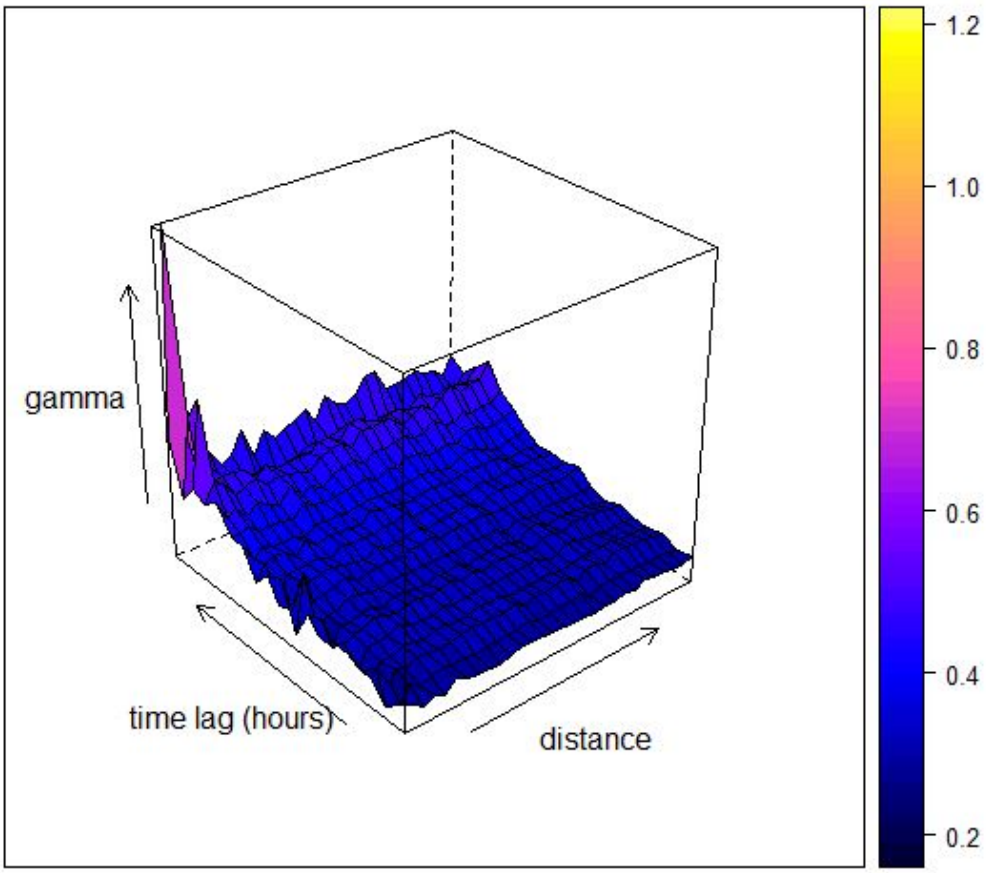

Figure 3.5: 시공간 배리오그램

3.4. 모수 배리오그램-공분산 모형 적합 (Parametric variogram-covariance model fitting)

이 절에서는 흔히 사용되는 모수 배리오그램-공분산 모형 (Parametric variogram-covariance model)을 몇 가지 소개하고 그 중에서 지수 모형 (Exponential model)을 지하철 데이터에 적합해보도록 한다.

(1) 구 모형 (Spherical model) :

$$\gamma_{a,b}^{sph}(\boldsymbol{h}) := \begin{cases} b\left(\frac{3}{2}\frac{\|\boldsymbol{h}\|}{a} - \frac{1}{2}\left(\frac{\|\boldsymbol{h}\|}{a}\right)^3\right), & \text{if } 0 \le \|\boldsymbol{h}\| \le a \\ b, & \text{otherwise} \end{cases}$$

$$C_{a,b}^{exp}(\boldsymbol{h}) := \begin{cases} b\left(1 - \frac{3}{2}\frac{\|\boldsymbol{h}\|}{a} + \frac{1}{2}\left(\frac{\|\boldsymbol{h}\|}{a}\right)^3\right), & \text{if } 0 \le \|\boldsymbol{h}\| \le a \\ 0, & \text{otherwise} \end{cases}$$

(2) 지수 모형 (Exponential model) :

$$\gamma_{a,b}^{exp}(\boldsymbol{h}) := b\left(1 - exp\big(-\frac{\|\boldsymbol{h}\|}{a}\big)\right), \quad \text{for } \|\boldsymbol{h}\| \ge 0$$

$$C_{a,b}^{exp}(\boldsymbol{h}) := bexp\big(-\frac{\|\boldsymbol{h}\|}{a}\big), \quad \text{for } \|\boldsymbol{h}\| \ge 0$$

(3) 가우시안 모형 (Gaussian model) :

$$\gamma_{a,b}^{gau}(\boldsymbol{h}) := b\left(1 - exp\big(-\frac{\|\boldsymbol{h}\|^2}{a^2}\big)\right), \quad \text{for } \|\boldsymbol{h}\| \ge 0$$

$$C_{a,b}^{gau}(\boldsymbol{h}) := bexp\big(-\frac{\|\boldsymbol{h}\|^2}{a^2}\big), \quad \text{for } \|\boldsymbol{h}\| \ge 0$$

(4) 매턴 모형 (Matern model) :

$$\gamma_{a,b,\nu}^{mat}(\boldsymbol{h}) := b\left[1 - \frac{1}{2^{\nu-1}\Gamma(\nu)}\left(\frac{\|\boldsymbol{h}\|}{a}\right)^{\nu} K_\nu\left(\frac{\|\boldsymbol{h}\|}{a}\right)\right], \quad \text{for } \|\boldsymbol{h}\| \ge 0$$

$$C_{a,b,\nu}^{mat}(\boldsymbol{h}) := b\left[1 - \frac{1}{2^{\nu-1}\Gamma(\nu)}\left(\frac{\|\boldsymbol{h}\|}{a}\right)^{\nu} K_\nu\left(\frac{\|\boldsymbol{h}\|}{a}\right)\right], \quad \text{for } \|\boldsymbol{h}\| \ge 0$$

매턴모형에서 $\Gamma(\cdot)$은 감마함수(Gamma function)이고 $K_\nu(\cdot)$은 수정된 베셀함수(modified Bessel function)이다. $\nu = 1/2$일 때 매턴모형은 지수모형과 동일하고 $\nu \to \infty$인 경우 가우시안 모형과 동일하다. 아래 그림은 지하철 데이터에 지수모형을 적합한 결과를 보여준다.

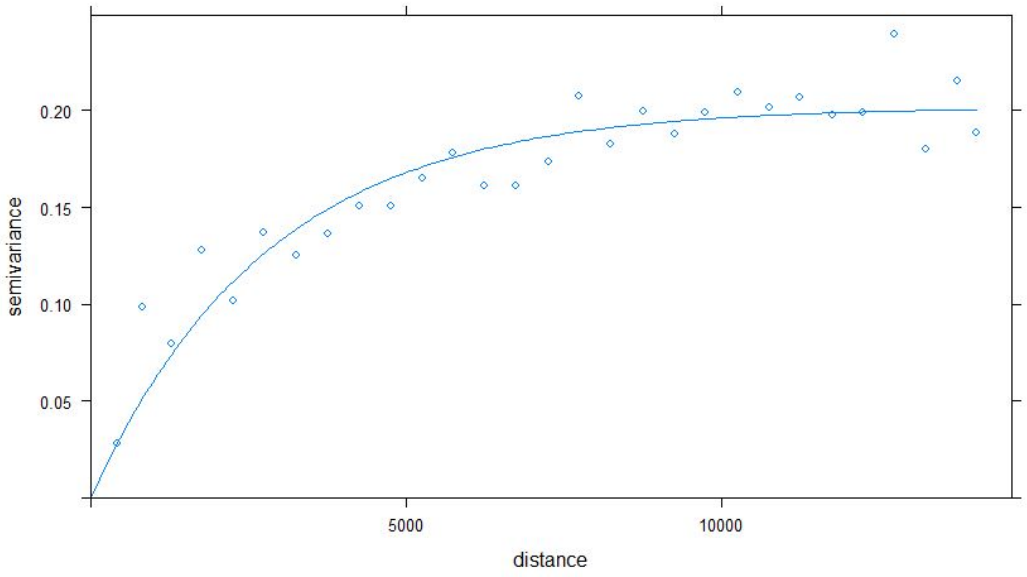

Figure 3.6: 8시 승차인원 지수 모형 배리오그램, $a = 2799.268$, $b = 0.2018$

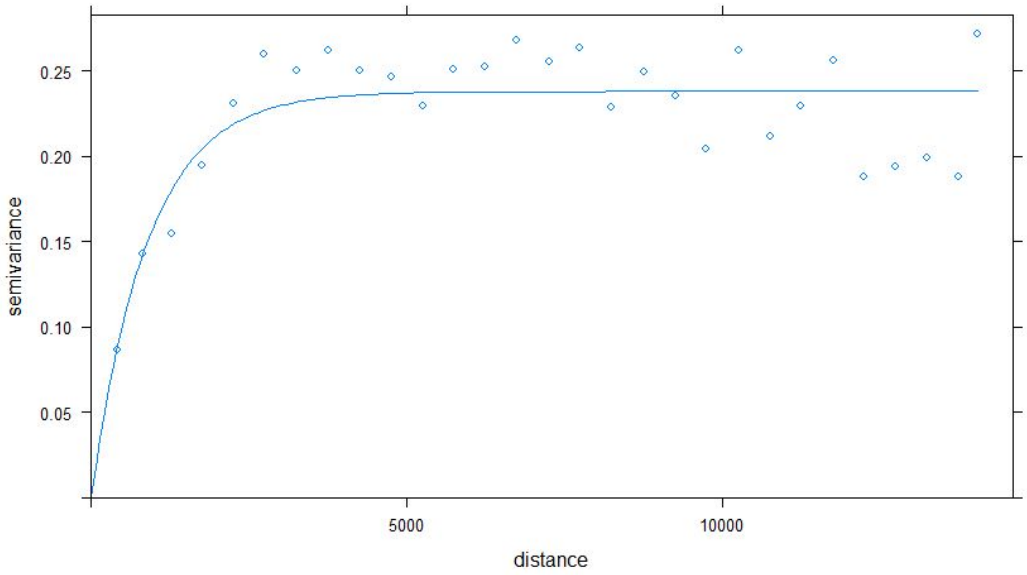

Figure 3.7: 18시 승차인원 지수 모형 배리오그램, $a = 897.6121$, $b = 0.2379$

Chapter 4

시공간 크리깅모형 (Spatio-Temporal Kriging Model)

4.1. 시공간 가우시안 과정 (Gaussian Process)

우선 147개의 지하철역 중 126개를 훈련자료 (Training Set), 21개를 시험자료 (Test Set) 로 나누자. 훈련자료 126개 역의 각 시점 (6시, 7시 ..., 24시) 에서 승차인원을 벡터로 나타내면 다음과 같다.

$$Y_o = \{y_{(stk)}\}, \quad s = 1, \ldots, 126, \, t = 1, \ldots, 19, \, k = 1, \ldots, 25$$

$$Y_o = (y_{(111)}, \, y_{(121)}, \, \cdots, \, y_{(1,19,1)}, \, y_{(112)}, \, \cdots, \, y_{(1,19,2)}, \, \cdots, \, y_{(126,19,25)})^t$$

이 때 아래첨자의 세 번째 값은 반복 (repetition) 을 나타내는데 우리가 가진 데이터는 2015년 1월부터 2017년 1월까지 한 달 단위로 측정된 데이터이므로 이를 반복으로 보면 k는 25까지의 값을 가지게 된다. 이와 마찬가지로 시험자료의 21개 지하철역 승차인원도 다음과 같이 벡터로 표현된다.

$$Y_p = \{y*_{(s'tk)}\}, \quad s' = 1, \ldots, 21,\ t = 1, \ldots, 19,\ k = 1, \ldots, 25$$

$$Y_p = (y*_{(111)},\ y*_{(121)},\ \ldots,\ y*_{(1,19,1)},\ y*_{(112)},\ \ldots,\ y*_{(1,19,2)},\ \ldots,\ y*_{(21,19,25)})^t$$

관측값 Y_o와 예측할 대상이 되는 값 Y_p의 결합분포를 다변량 정규분포로 가정하면 아래와 같이 쓸 수 있다.

$$Y = \begin{pmatrix} Y_o \\ Y_p \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_o \\ \mu_p \end{pmatrix}, \sigma^2 \begin{pmatrix} \Sigma_o & \Sigma_{op} \\ \Sigma_{po} & \Sigma_p \end{pmatrix} \right] \tag{4.1}$$

벡터 Y의 총길이는 $(126 + 21) \times 19 \times 25 = 69825$가 되며 매월 승차인원은 서로 독립이므로 Σ_o는 다음과 같은 블럭대각행렬(Block diagonal matrix)의 형태가 된다.

$$\Sigma_o = \begin{bmatrix} \Sigma_{st} & \mathbf{0}_{126\times19} & \cdots & \mathbf{0}_{126\times19} \\ \mathbf{0}_{126\times19} & \Sigma_{st} & \cdots & \mathbf{0}_{126\times19} \\ \vdots & & \ddots & \vdots \\ \mathbf{0}_{126\times19} & \cdots & \mathbf{0}_{126\times19} & \Sigma_{st} \end{bmatrix}$$

다음 절에서 평균 구조(Mean Structure)인 $\begin{pmatrix} \mu_o \\ \mu_p \end{pmatrix}$와 공분산 구조(Covariance Structure)인 $\begin{pmatrix} \Sigma_o & \Sigma_{op} \\ \Sigma_{po} & \Sigma_p \end{pmatrix}$에 대해 설명한다.

4.2. 평균 구조(Mean Structure)

지하철역의 시간에 따른 승차인원 Y와 이에 영향을 주는 공변량(Covariate)으로 환승역 여부($\boldsymbol{x}_1$: 0 or 1), 계절($\boldsymbol{x}_2$: 1 or 2 or 3 or 4), 근처 버스 정류장의 시간에 따른 승차인원($\boldsymbol{x}_3$) 등을 고려하여 다음과 같은 시공간 회귀 모형

(Spatio-temporal regression model)을 생각하자.

$$Y^{stk} = \beta_0 + \beta_1 \boldsymbol{x}_1^s + \beta_2 \boldsymbol{x}_2^k + \beta_3 \boldsymbol{x}_3^{stk} + \varepsilon^{st}$$

여기에 시간에 따른 효과를 잡아주기 위한 B-스플라인(B-spline) 곡선을 추가한 다음과 같은 모형을 고려하자. 이 때 B-스플라인 곡선의 차수(degree)는 3으로 하고 내부 마디(Internal knot)의 수는 4개로 하였다. 즉 기저함수(Basis function)의 수는 7개이다.

$$Y^{stk} = \beta_0 + \beta_1 \boldsymbol{x}_1^s + \beta_2 \boldsymbol{x}_2^k + \beta_3 \boldsymbol{x}_3^{stk} + \sum_{j=1}^{7} b_j B_j(t) + \varepsilon^{st}$$

이와 같은 회귀모형을 고려하면 식 (4.1)의 평균벡터 $\begin{pmatrix} \mu_o \\ \mu_p \end{pmatrix}$는 설계행렬(Design Matrix)과 추정이 필요한 모수벡터 $\boldsymbol{\beta}$의 곱으로 나타내어진다.

$$\begin{pmatrix} \mu_o \\ \mu_p \end{pmatrix} = \begin{pmatrix} X_o \boldsymbol{\beta} \\ X_p \boldsymbol{\beta} \end{pmatrix}$$

$\boldsymbol{x}_1^s$과 $\boldsymbol{x}_2^k$가 각각 2가지 범주와 4가지 범주를 가지는 범주형 변수이므로 관측치에 대한 설계행렬 X_o의 크기는 $(126 * 19 * 25) \times 13$가 된다.

두 가지 범주형 변수인 $\boldsymbol{x}_1^s$과 $\boldsymbol{x}_2^k$를 고려한 설계행렬을 구성하기 위해 R의 함수 `model.matrix()`를 사용하였다.

4.3. 공분산 구조(Covariance Structure)

이번 절에서는 식 (4.1)의 공분산 행렬의 구조에 대해 살펴보도록 한다. 모형의 간소화를 위해 시간에 관한 의존성과 공간에 관한 의존성이 분리가능(Separable)하다고 하자. 즉, 공간에 관한 의존성이 시간에 따라 달라지지 않음을 의미한다. 시공간 배리오그램(그림 3.5)을 통해서도 분리가능성 가정이

타당함을 볼 수 있다. 이 경우 공분산 행렬 Σ_{st}은 공간 공분산 행렬(Spatial covariance matrix) Σ_s와 시간 공분산 행렬(Temporal covariance matrix) Σ_t의 크로네커곱(Kronecker product)으로 나타내어진다. 식으로 표현하면

$$\Sigma_{st} = \Sigma_s \otimes \Sigma_t$$

이며, Σ_s와 Σ_t는 지수 공분한 함수(Exponential covariance function)를 적용하여 다음과 같다.

$$\Sigma_s = \begin{cases} e^{-\rho_s dist(s_i, s_j)} & \text{for } i \neq j \\ 1 & \text{for } i = j \end{cases}$$

$$\Sigma_t = \begin{cases} e^{-\rho_t |t_{i'} - t_{j'}|} & \text{for } i' \neq j' \\ 1 & \text{for } i' = j' \end{cases}$$

Σ_s는 두 지점의 거리 $dist(s_i, s_j)$에 따라 다른 값을 가지고 Σ_t는 시차(lag) $| t_{i'} - t_{j'} |$에 따라 다른 값을 가진다. 이와 같은 공분산행렬의 분리가능성으로 인해 모수의 최대우도추정값(Maximum Likelihood Estimator)을 더 쉽게 구할 수 있다.

4.4. 최대우도추정(Maximum Likelihood Estimator)

앞에서 정규분포를 가정하고 평균구조, 공분산구조에 대해 정의해줌으로써 모수의 최대우도추정(Maximum Likelihood Estimate)에 대하여 다음과 같은 전개가 가능하게 되었다. 먼저 X_o와 Y_o를 반복에 따라 분리하여 표현하면

다음과 같다.

$$X_o = \begin{pmatrix} X_o^{..1} \\ X_o^{..2} \\ \vdots \\ X_o^{..25} \end{pmatrix}$$

$$Y_o = \begin{pmatrix} Y_o^{..1} \\ Y_o^{..2} \\ \vdots \\ Y_o^{..25} \end{pmatrix}$$

우리가 추정해야 하는 모수의 집합은 $\theta = (\boldsymbol{\beta}, \sigma^2, \rho_s, \rho_t)$ 인데 Σ_o의 블럭대각행렬 구조와 위와 같이 표현된 X_o와 Y_o의 구조를 이용하면 $\boldsymbol{\beta}$의 최대우도추정량은 다음과 같이 표현된다.

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (X_o^T \Sigma_o^{-1} X_o)^{-1} X_o^T \Sigma_o^{-1} Y_o \\ &= \sum_{k=1}^{25} (X_o^{..k\,T} \Sigma_{st}^{-1} X_o^{..k})^{-1} X_o^{..k\,T} \Sigma_{st} Y_o^{..k} \\ &= \sum_{k=1}^{25} (X_o^{..k\,T} [\Sigma_s^{-1} \otimes \Sigma_t^{-1}] X_o^{..k})^{-1} X_o^{..k\,T} [\Sigma_s \otimes \Sigma_t] Y_o^{..k} \end{aligned} \tag{4.2}$$

다음으로 σ^2의 최대우도추정량은

$$\begin{aligned} \hat{\sigma^2} &= \frac{1}{N} (Y_o - X_o \hat{\boldsymbol{\beta}})^T \Sigma_o^{-1} (Y_o - X_o \hat{\boldsymbol{\beta}}) \\ &= \frac{1}{N} \sum_{k=1}^{25} (Y_o^{..k} - X_o^{..k} \hat{\boldsymbol{\beta}})^T \Sigma_{st}^{-1} (Y_o^{..k} - X_o^{..k} \hat{\boldsymbol{\beta}}) \\ &= \frac{1}{N} \sum_{k=1}^{25} (Y_o^{..k} - X_o^{..k} \hat{\boldsymbol{\beta}})^T [\Sigma_s^{-1} \otimes \Sigma_t^{-1}] (Y_o^{..k} - X_o^{..k} \hat{\boldsymbol{\beta}}) \end{aligned} \tag{4.3}$$

이다. 여기서 N은 관측데이터의 총 수인 $126 \times 19 \times 25 = 59850$ 이다. 이제 $\hat{\rho}_s$과 $\hat{\rho}_t$를 구하기 위해 우도함수(Likelihood function)를 살펴보자. 우도함수를 L이라고 하면

$$-2\log L \propto \log|\ \sigma^2\Sigma_o\ | + \frac{1}{\sigma^2}(Y_o - X_o\boldsymbol{\beta})^T\Sigma_o^{-1}(Y_o - X_o\boldsymbol{\beta})$$

이다. 여기에 $\hat{\boldsymbol{\beta}}$와 $\hat{\sigma^2}$를 대입하여 전개하면

$$\begin{aligned} -2\log L \propto & \log|\ \hat{\sigma^2}\Sigma_o\ | + \frac{1}{\hat{\sigma^2}}(Y_o - X_o\hat{\boldsymbol{\beta}})^T\Sigma_o^{-1}(Y_o - X_o\hat{\boldsymbol{\beta}}) \\ &= \log|\ \hat{\sigma^2}\Sigma_o\ | + N \\ &= \log \hat{\sigma^2}^N + \log\left(|\ \Sigma_s\ |^{19}|\ \Sigma_t\ |^{126}\right)^{25} \\ &= N\log\hat{\sigma^2} + 25\cdot 19\log|\ \Sigma_s\ | + 25\cdot 126\log|\ \Sigma_t\ | \end{aligned} \tag{4.4}$$

가 된다. 위의 식을 최소화 하는 ρ_s와 ρ_t를 수치적 방법으로 구하여 최종적으로 $\hat{\Sigma_s}$,와 $\hat{\boldsymbol{\theta}}^{MLE}$를 얻게 된다.

4.5. 일반 크리깅 (Universal Kriging)

Y_o와 Y_p의 결합확률분포를 정규분포로 가정하였으므로

$$Y_p\ |\ Y_o \sim N\Big[X_p\boldsymbol{\beta} + \Sigma_{po}\Sigma_o^{-1}(Y_o - X_o\boldsymbol{\beta}),\ \sigma^2(\Sigma_p - \Sigma_{po}\Sigma_o^{-1}\Sigma_{op})\Big]$$

가 된다. 이 때 조건부 기댓값 $E(Y_p\ |\ Y_o) = X_p\boldsymbol{\beta} + \Sigma_{po}\Sigma_o^{-1}(Y_o - X_o\boldsymbol{\beta})$은 Y_p의 최적선형불편예측값(Best Linear Unbiased Predictor)이 된다. 따라서 Y_p의 예측값(Prediction)은

$$\hat{Y_p} = X_p\hat{\boldsymbol{\beta}} + \hat{\Sigma}_{po}\hat{\Sigma}_o^{-1}\big(Y_o - X_o\hat{\boldsymbol{\beta}}\big) \tag{4.5}$$

가 된다. 또한 평균제곱예측오차(Mean Square Prediction Error)

$$E\Big[\big(Y_p - g(Y_o)\big)^2\Big]$$

를 가장 작게 하는 g를 생각하여도

$$g(Y_o) = E(Y_p \mid Y_o) = X_p\boldsymbol{\beta} + \Sigma_{po}\Sigma_o^{-1}(Y_o - X_o\boldsymbol{\beta})$$

는 최적예측값(Best Predictor)이 된다. 조건부 분산행렬 $V(Y_p \mid Y_o) = \sigma^2\big(\Sigma_p - \Sigma_{po}\Sigma_o^{-1}\Sigma_{op}\big)$의 추정을 통해 예측구간(Prediction interval)을 구할 수도 있다.

$$\begin{aligned} \hat{V}(Y_p \mid Y_o) &= \hat{\sigma^2}\big(\hat{\Sigma}_p - \hat{\Sigma}_{po}\hat{\Sigma}_o^{-1}\hat{\Sigma}_{op}\big) \\ \tilde{\sigma_p} &= \sqrt{diag\big(\hat{V}(Y_p \mid Y_o)\big)} \end{aligned} \tag{4.6}$$

라고 하면 95% 예측구간은 $\hat{Y_p} \pm 1.96\tilde{\sigma_p}$가 된다.

Chapter 5

데이터 분석(Data Analysis)

이 장에서는 앞의 4장에서 설명한 내용을 지하철 데이터에 적용한 결과를 기술한다. 먼저 중부원점 좌표계로 표시된 위치정보를 편의상 1/1000로 스케일링하여 분석을 진행하였고 분석에는 version 3.4.0 R이 사용되었다.

5.1. 예측 결과

R의 `optim()` 함수를 이용하여 식 (4.4)를 최소화 하는 ρ_s와 ρ_t를 구하면 각각 0.596과 0.0223이다. 이를 통해 $\hat{\Sigma_s}$와 $\hat{\Sigma_t}$를 구할 수 있고 이를 식 (4.2)와 (4.3)에 대입하여 구한 $\hat{\boldsymbol{\beta}}$와 $\hat{\sigma^2}$는 다음과 같다.

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} 3.9527 \\ 0.4831 \\ 0.066 \\ 0.0458 \\ 0.0391 \\ 0.7377 \\ 0.0851 \\ 0.2014 \\ 0.5244 \\ 0.0511 \\ 0.2232 \\ -1.2122 \\ 3.34 \times 10^{-6} \end{pmatrix}$$

$$\hat{\sigma^2} = 0.2792$$

$\hat{\boldsymbol{\beta}} = (\hat{\beta_0}, \hat{\beta_1}, \ldots, \hat{\beta_{12}})$라고 할 때, $\hat{\beta_1} = 0.4831$은 환승역 여부에 대한 효과, $(\hat{\beta_2}, \hat{\beta_3}, \hat{\beta_4}) = (0.066, 0.0458, 0.0391)$은 계절에 대한 효과, $(\hat{\beta_5}, \ldots, \hat{\beta_{11}}) = (0.7377, \ldots, -1.2122)$는 B-스플라인 곡선 계수, 마지막으로 $\hat{\beta}_{12} = 3.34 \times 10^{-6}$은 주변 버스정류장의 승차인원에 대한 효과를 나타낸다. $\hat{\beta}_{12}$값이 아주 작게 나온 이유는 Y에 해당하는 지하철역 승차인원은 로그변환을 해준 반면 버스 승차인원은 원래의 값을 그대로 사용하였기 때문이다.
$\hat{\boldsymbol{\beta}}$의 분포가 $N\left[\boldsymbol{\beta},\, \sigma^2\left(X_o^T \Sigma_o X_o\right)^{-1}\right]$ 임을 이용하여 구한 $\hat{\boldsymbol{\beta}}$의 95% 신뢰구간을 Tabe 5.1에서 볼 수 있다.

$\boldsymbol{\beta}$	추정값	95% 신뢰구간 하한	95% 신뢰구간 상한
β_0	3.9527	3.877	4.0285
β_1	0.4831	0.455	0.5113
β_2	0.066	-0.0423	0.1743
β_3	0.0458	-0.0625	0.1541
β_4	0.0391	-0.0692	0.1474
β_5	0.7377	0.7218	0.7536
β_6	0.0851	0.0615	0.1086
β_7	0.2014	0.1739	0.2288
β_8	0.5244	0.4928	0.5561
β_9	0.0511	0.0154	0.0869
β_{10}	0.2232	0.1873	0.2591
β_{11}	-1.2122	-1.2469	-1.1775
β_{12}	3.3482×10^{-6}	3.1549×10^{-6}	3.5415×10^{-6}

Table 5.1: $\boldsymbol{\beta}$ 추정값 95% 신뢰구간

위에서 구한 $\hat{\boldsymbol{\beta}}$ 값을 식 (4.5)와 (4.6)에 대입하여 시험 자료 21개 지하철역에 대한 예측값과 예측구간을 구할 수 있다. 시험 자료로 분류된 21개 지하철역 중 5개 역의 2015년 1월 승차인원 예측값과 실제값을 비교한 그림을 보면 예측값이 실제값을 대체적으로 잘 따라감을 볼 수 있다.

실제로 우리는 총 $21 \times 19 \times 25 = 9975$개의 예측값과 예측구간을 가지게 되는데, 9,975개의 예측구간 중 실제 승차인원의 값을 포함하는 것은 총 9,117개로 전체의 91.4%이다.

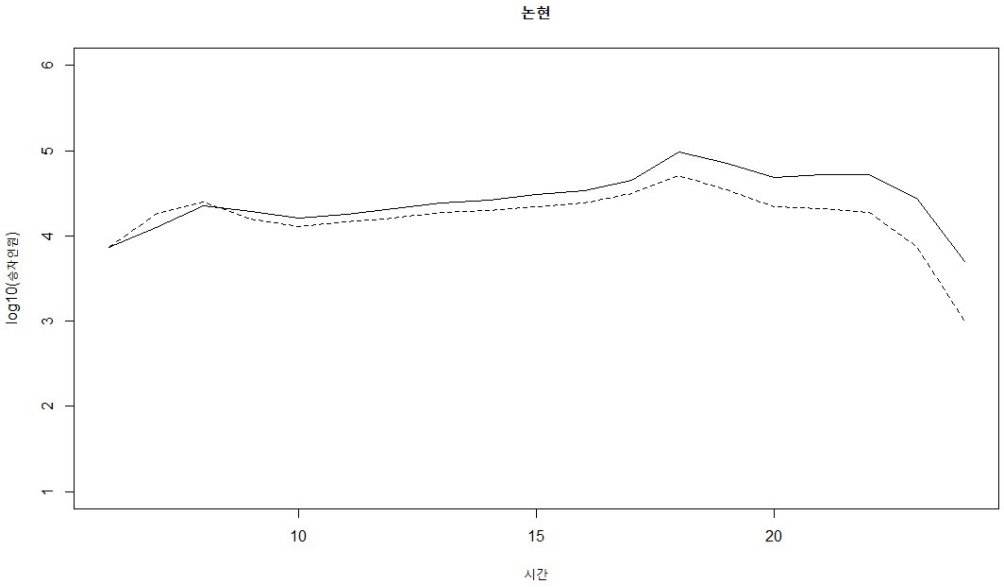

Figure 5.1: 논현역 예측값(점선)과 실제값(실선)

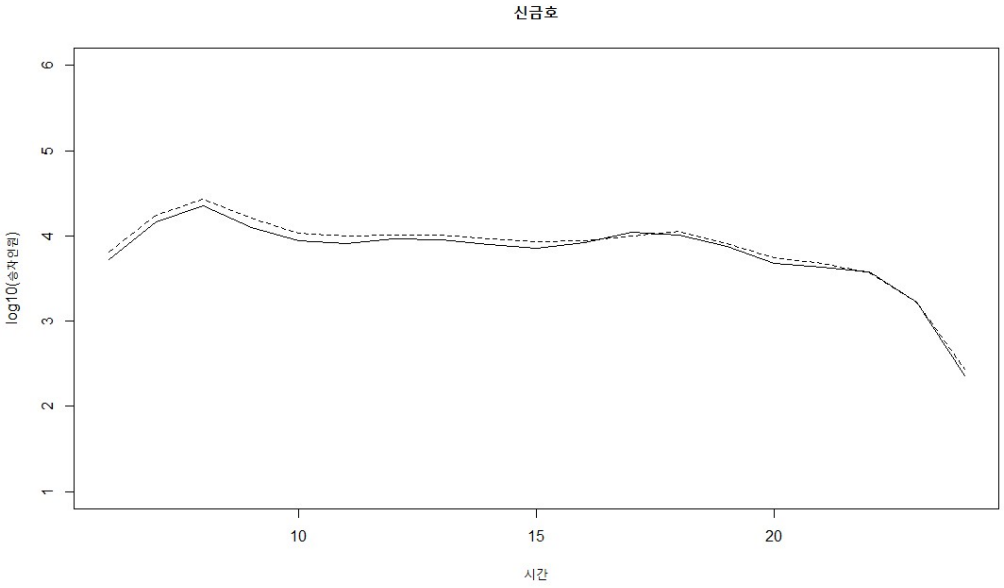

Figure 5.2: 신금호역 예측값(점선)과 실제값(실선)

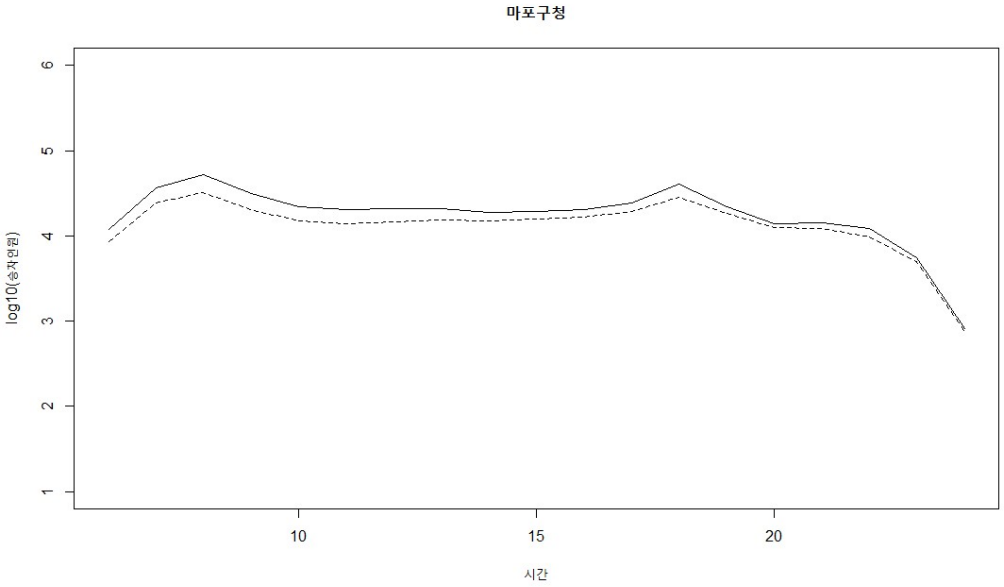

Figure 5.3: 마포구청역 예측값(점선)과 실제값(실선)

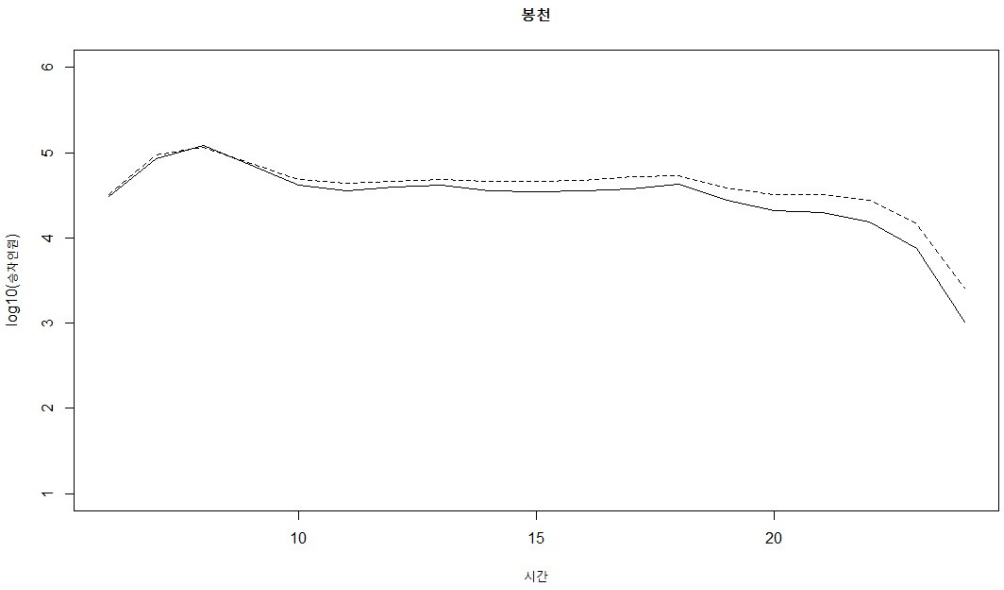

Figure 5.4: 봉천역 예측값(점선)과 실제값(실선)

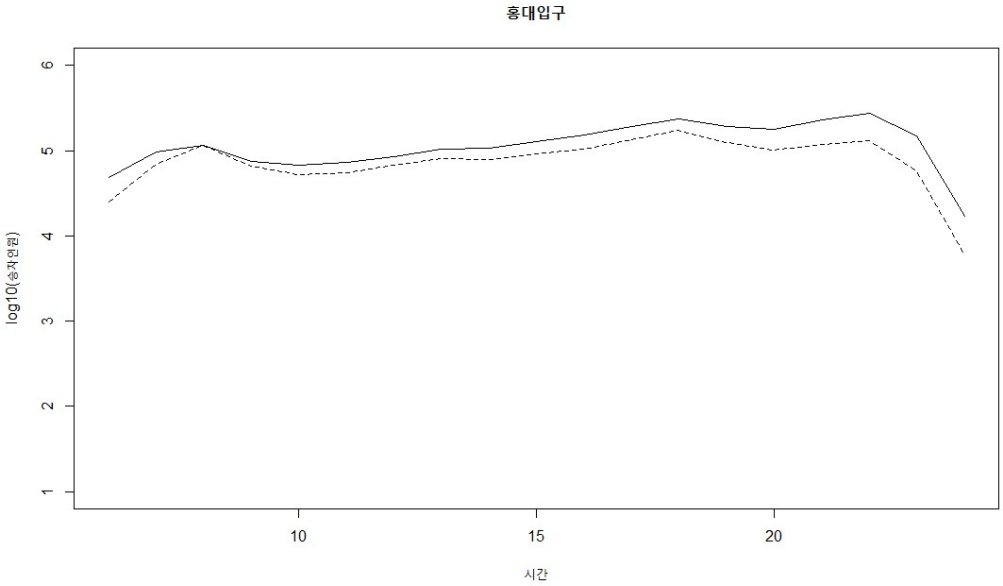

Figure 5.5: 홍대입구역 예측값(점선)과 실제값(실선)

Chapter 6

결론(Conclusion)

지금까지 지하철 승차인원 예측을 위한 시공간 모형화(Spatio-teporal modelling)와 이에 대한 이론적 배경을 설명하였다. 3장에서 데이터의 시간, 공간적 의존성을 살피는데 유용한 도구가 되는 배리오그램(Variogram)으로 지하철 데이터의 의존성(Dependency)을 탐색하였고 4장에서는 시공간 가우시안 모형(Spatio-temporal Gaussian model)을 구성하는 평균구조(Mean structure)와 공분산구조(Covariance structure)에 대해 설명하였다. 모형의 간소화를 위해 시간과 공간에 대한 의존성이 분리가능(Separable)하다고 가정하였고 모형에서 Y에 해당하는 지하철 승차인원에 영향을 끼칠만한 공변량(Covariates)을 몇 가지 선정하고 또 B-스플라인(B-spline)인 곡선을 추가하여 설계행렬을 구성하였다. 마지막으로 5장에서는 모형의 예측 성능에 대해 살펴보았다. 본 논문을 결론짓기에 앞서 우리의 시공간 모형으로 서론에서 언급한 신림선(샛강 ~ 서울대)의 승차인원에 대한 예측을 해보도록 하자. 그림 6.1은 신림선을 구성하는 10개 역의 위치를 보여준다.

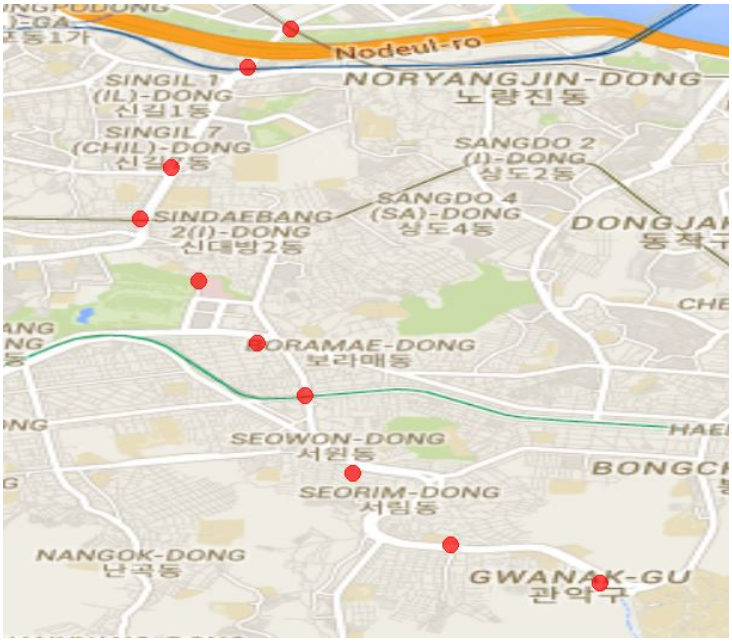

Figure 6.1: 새로 개통될 신림선의 위치

신림선을 구성하는 10개 역의 연간 승차인원 예측값이 표 6.1에 나타나 있다. 괄호 속의 값은 기존에 이미 있던 역인 샛강, 대방, 보라매, 신림의 2016년 승차인원값을 보여준다. 더 정확한 예측을 위해 모형을 발전시킬 수 있는데 우선 월별로 나누어진 자료를 반복으로 보지 않고 이에 대한 의존성을 고려한 모형을 생각할 수 있으며 승차인원뿐 아니라 시간대별 하차인원도 고려하여 승하차간의 의존성이 표현된 모형도 생각해 볼 수 있다.

역 이름	예측값
샛강	2384047 (1831699)
대방	3908465 (5727695)
병무청	3808764
보라매	3628676 (3440407)
보라매병원	1934860
당곡	3369247
신림	6722320 (26133552)
서원	4573427
서림	3346569
서울대	6413828

Table 6.1: 신림선의 연간 승차인원 예측값

참고문헌

[1] Cressie, N. (2015). Statistics for spatial data. John Wiley & Sons.

[2] Racine, J. S. (2014). A primer on regression splines.

[3] Lichtenstern, A. (2013). Kriging methods in spatial statistics. Technische Universität München.

[4] Wackernagel, H. (2013). Multivariate geostatistics: an introduction with applications. Springer Science & Business Media.

[5] Ciampalini, R., Lagacherie, P., Monestiez, P., Walker, E., & Gomez, C. (2012). Co-kriging of soil properties with Vis-NIR hyperspectral covariates in the Cap Bon region (Tunisia). In Digital Soil Assessments and Beyond (pp. 393-398). CRC Press.

[6] Schelin, L. (2012). Spatial sampling and prediction (Doctoral dissertation, Umeå universitet).

[7] Agarwal, A. (2011). A New Approach to Spatio-temporal Kriging and Its Applications (Doctoral dissertation, The Ohio State University).

[8] Gottfridsson, A. (2011). Likelihood ratio tests of separable or double separable covariance structure, and the empirical null distribution.

[9] Renard, D. (2011). Roger S. Bivand, Edzer J. Pebesma, Virgilio Gomez-Rubio: Applied Spatial Data Analysis with R. Mathematical Geosciences, 43(5), 607-609.

[10] Pebesma, E. J. (2004). Multivariable geostatistics in S: the gstat package. Computers & Geosciences, 30(7), 683-691.

[11] Zimmerman, D. L., & Cressie, N. (1992). Mean squared prediction error in the spatial linear model with estimated covariance parameters. Annals of the institute of statistical mathematics, 44(1), 27-43.

Abstract

Minwoo Kim
The Department of Statistics
Graduate School
Seoul National University

The subway is a mode of transportation used by many people because of its quickness and convenience. This paper is written to predict subway passenger flows by using the statistical modeling. Among many types of data, spatial data which includes locational information is now gaining more attraction and statistical methodology considering spatial dependency has been developed. Each subway station in Seoul has been assigned locational information and the spatio-temporal model which considers temporal dependency as well as spational dependency is used to predict subway passenger flows. The subway passenger flows are assumed to be normally distributed and mean structure consists of several covariates and B-spline curves. Separability between temporal dependency and spatial dependency is assumed and exponential model is used in the covariance structure. Parameters for the final model are estimated by the maximum likelihood method.

Keyword :*Spatial data, Spatial-temporal model, B-spline, Separability, Maximum Likelihood Estimates*
Student Number : 2015-20289

국문초록

시공간 모형화를 통한
지하철역 승차인원 예측
Predicting Subway Passenger Flows
By Spatio-temporal Modeling

지하철은 신속성과 편리함으로 인해 많은 사람들이 이용하는 교통수단이다. 통계적 모형화를 통해 시간대별 지하철 승차인원을 예측하고자 이 논문이 쓰여졌다. 많은 종류의 데이터 중에서 위치정보를 포함한 공간 데이터가 주목을 받고 있으며 위치정보에 의한 의존성을 고려한 공간 통계 방법론이 발전하고 있다. 서울의 각 지하철역에 위치정보를 부여하고 공간적 의존성과 시간에 따른 의존성을 함께 고려한 시공간 모형화를 통해 지하철 승차인원을 예측하려 한다. 지하철 승차인원을 정규분포로 가정하고 몇 가지 공변량과 B-스플라인 곡선으로 평균구조를 구성한다. 공분산구조는 시간에 따른 의존성과 공간에 의한 의존성이 분리가능하다고 가정하여 지수모형을 적용하며 최종 모형의 모수는 최대우도 추정법으로 추정된다.

주요어 : 공간 데이터, 시공간 모형, B-스플라인, 분리가능, 최대우도 추정

학 번 : 2015-20289