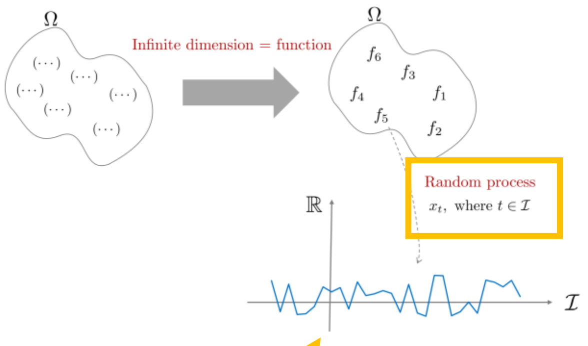
Spatial-temporal modeling particulate matter concentration

이민주

EWHA,
THE FUTURE
WE CREATE

SPDE

Geostatistical Data



- Geostatistical Data의 한계 → 점의 정보이기 때문에 빈 공간의 정보를 알기 힘듦.
- 몇 가지 가정(**Isotropic, Stationary**)를 활용하여 빈 공간을 채울 때(=Kriging) 도움을 받고자 함.

$$\{Z(s): s \in D \subset \mathbb{R}^2\}$$

:주어진 도메인(D, 예: 서울시) 내에 내포된 위치 s에서의 관측값들

$s_1, \dots, s_n$: 위치 $Z(s_1), \dots, Z(s_n)$: 위치 s에서의 무작위 과정(Random Process)

- **Weak Stationary:** $\mathrm{E}[\mathrm{Z}(\mathbf{s})] = \mu$, $\mathrm{Cov}(\mathrm{Z}(\mathbf{s_1}, \mathbf{s_2})) = \mathrm{K}(\mathbf{s_1} - \mathbf{s_2})$, $\mathbf{s_1} - \mathbf{s_2} \neq \mathbf{s_2} - \mathbf{s_1}$

방향 & 거리 동일하면 Covariance 같음.

- Strong Stationary: $\{Z(\boldsymbol{s_1}), \dots, Z(\boldsymbol{s_n})\} = \{Z(\boldsymbol{s_1} + \boldsymbol{h}), \dots, Z(\boldsymbol{s_n} + \boldsymbol{h})\}$ → 너무 까다로운 조건, 현실성x

- **Weak Isotropic:** $\mathrm{Cov}(\mathrm{Z}(\mathbf{s_1}, \mathbf{s_2})) = \mathrm{K}(|\mathbf{s_1} - \mathbf{s_2}|)$, $\mathbf{s_1} - \mathbf{s_2} = \mathbf{s_2} - \mathbf{s_1}$

방향 고려x, 거리만 동일하면 Covariance 같음.

Autocovariance(자기공분산)

: s_i에 h(거리)를 더함으로써 covariance를 구할 수 있음

THE FUTURE WE CREATE

Matérn Covariance Function

- **Matérn Covariance Function:** 공간 통계학에서 자주 활용되는 Autocovariance 함수. **장점: 공간적으로 가까운 지점들의 유사성 파악함** SPDE는 Matérn Covariance Function 으로 설명되는 Process 계산 방법으로 활용.
- **SPDE(Stochastic Partial Differentials Equation):** 유현하고 효율적인 방법으로 geostatistical data 모델링 및 예측하는 방법론. **장점: 연속적인 공간 Process 분포를 얻어 샘플링되지 않은 위치 예측 가능.**

SPDE

Gaussian Spatial White Noise

$$(\kappa^2 - \Delta)^{\frac{\alpha}{2}}\left(\tau\underbrace{\xi(\boldsymbol{s})}_{\text{Stationary GF (Gaussian Field)}}\right) = \boxed{\mathcal{W}(s)}$$

Matérn Covariance Function

Bassel Function (2nd order): 매끄러움 측정

$$\mathrm{Cov}\left(\xi_i(\boldsymbol{s_i}), \xi_j(\boldsymbol{s_j})\right) = \mathrm{Cov}(\xi_i, \xi_j) = \frac{1}{\Gamma(\lambda)2^{\lambda-1}}\left(\kappa\underbrace{\|s_i - s_j\|}\right)^{\lambda}\boxed{K_\lambda}\left(\kappa\underbrace{\|s_i - s_j\|}\right)$$

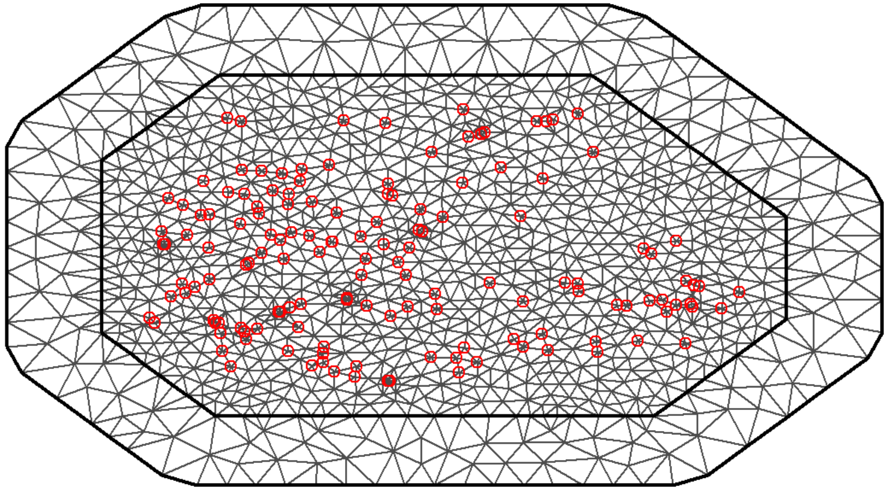
지역 i와 j의 Euclidean Distance

$$\begin{cases} \alpha: \text{smoothness} \\ \kappa > 0: \text{scale parameter} \\ \tau: \text{variance} \end{cases}$$

- Hyperparameter: $\boldsymbol{\psi} = \{\sigma_e^2, \kappa, \sigma^2\}$
- Parameter: $\boldsymbol{\theta} = \{\tilde{\boldsymbol{\xi}}, b_0\}$

EWHA, THE FUTURE WE CREATE

Matérn Covariance Function



- $\boldsymbol{s} \in \mathbb{R}^2$ (d(dimension) $= 2$), (디폴트) $\alpha = 2, \lambda = 1$ 일때,

$\rightarrow \begin{cases} r = \frac{\sqrt{8}}{\kappa} \\ \sigma^2 = \frac{1}{4\pi\kappa^2\tau^2} \end{cases}$:공간적 상관성이 있다고 여겨지는 기준(범위)

SPDE

$$(\kappa^2 - \Delta)^{\frac{\alpha}{2}}(\tau\xi(\boldsymbol{s})) = \mathcal{W}(s)$$

$\xi(\boldsymbol{s}) = \sum_{g=1}^{G} \varphi_g(\boldsymbol{s})\,\widetilde{\xi_g}$, g = mesh에서 삼각형 꼭짓점의 개수

$\varphi_g(\boldsymbol{s})$: g 꼭짓점에 위치하면 1, 아니면 0.

$\tilde{\xi} = \{\widetilde{\xi_1}, .., \widetilde{\xi_G}\}$: Gaussian Weight Vector

EWHA,
THE FUTURE
WE CREATE

Linear Prediction

https://www.paulamoraga.com/book-spatial/sec-geostatisticaldataSPDE.html

Linear Prediction

$$\eta_i = b_0 + \sum_{g=1}^{G} \varphi_g(\boldsymbol{s})\,\widetilde{\xi_g} = b_0 + \sum_{g=1}^{G} A_{ig}\,\widetilde{\xi_g}$$

Projection Matrix

nrow = n(obs 개수), ncol = G(mesh 삼각형 꼭짓점 개수)

Projection Matrix

$\sum_{g=1}^{G} A_{ig} = 1$

$$\boxed{A} = \begin{bmatrix} A_{11} & A_{12} & A_{13} & \dots & A_{1G} \\ A_{21} & A_{22} & A_{23} & \dots & A_{2G} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & A_{n3} & \dots & A_{nG} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & \dots & 0 \\ A_{21} & A_{22} & 0 & \dots & A_{2G} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & A_{n3} & \dots & 0 \end{bmatrix} \begin{array}{l} \rightarrow \text{1st observation} \\ \rightarrow \text{2nd observation} \\ \\ \\ \end{array}$$

3번째 observation은 3번째
꼭짓점에 정확히 위치함.

EWHA,
THE FUTURE
WE CREATE

Advanced Operation: inla.stack

: Model Complexity가 증가할수록 SPDE의 에러가 발생할 수 있음 → INLA.stack 활용

Linear Prediction

$$\eta_i = \sum_k \boxed{h_k}\boxed{(Z_i^k)}$$

공변량 value

Link function (각 공변량을 nonlinear effect, random effect 등 적절한 함수에 대입함)

Linear Algebra로 표현

$$\eta^* = \eta_1^* + \eta_2^* + \ldots = A_1 H(Z_1) + A_2 H(Z_2) + \ldots = [A_1, A_2, \ldots] H\left(\begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \end{bmatrix}\right) = \tilde{A} H(\tilde{Z})$$

Spatio-temporal modeling particulate matter concentration

Ⅰ. Air pollution in Piemonte Dataset

Data Explanation

- Geostatistical Data
- 2005년 10월 ~ 2006년 3월까지 (매일 측정, 총 182일) Piemonte 지역 (북부 이탈리아)의 PM_{10} 농도 분석
- 총 24개의 관측소(점)을 분석
- nrow: 4368 (24(station) x 182(days)), ncol: 12

Covariates X (M=8) : scaled

Y

Log(Y)

Piemonte Data

관측소: i=1,..,24

Station.ID	Date	A	UTMX	UTMY	WS	TEMP	HMIX	PREC	EMI	PM10	time	logPM10
1	01/10/05	-1.3955833	0.9401862	-0.6364774	-0.0776828	2.096996	2.181230	-0.2902373	-0.1752838	28	1	3.332205
2	01/10/05	-0.7563919	-0.2639052	-1.1771843	-0.2319185	2.068951	1.691691	-0.2902373	-0.3453852	22	1	3.091042
3	01/10/05	-0.0253574	1.4970490	-1.2218365	0.0379940	1.822551	2.527195	-0.2902373	-0.6353254	17	1	2.833213
4	01/10/05	-0.8881265	0.0996534	-0.6245213	0.4428627	2.060938	1.383675	-0.2902373	-0.0984938	25	1	3.218876
5	01/10/05	1.4784586	-0.1863740	1.1400961	-0.6560668	1.860613	1.371974	-0.2902373	-0.0324079	20	1	2.995732
6	01/10/05	0.1100881	-1.0203586	0.0547783	0.1536708	2.052925	1.402333	-0.2902373	0.0152950	41	1	3.713572

날짜: 10월 1일 2005년 ~
3월 31일 2006년 (총 182일)

X축, Y축 좌표계 (covariate에 포함됨)

Time(일): t=1,..,182

EWHA, THE FUTURE WE CREATE

Spatial-temporal Model

$y_{it} \sim Normal(\eta_{it}, \sigma_e^2)$ $\quad$ i = 지역(1,..,24), t=년도(1,…,182), m=공변량(1,..,8)

Linear Prediction

$$\eta_{it} = b_0 + \sum_{m=1}^{M} \beta_m x_{mi} + \omega_{it}$$

잠재적 spatio-temporal process

$$\omega_{it} = a\omega_{i(t-1)} + \xi_{it}$$

Zero-mean Gaussian Field

$$\mathrm{Cov}(\xi_{it}, \xi_{ju}) = \begin{cases} 0 & \text{if } t \neq u \\ \mathrm{Cov}(\xi_i, \xi_j) & \text{if } t = u \end{cases}$$

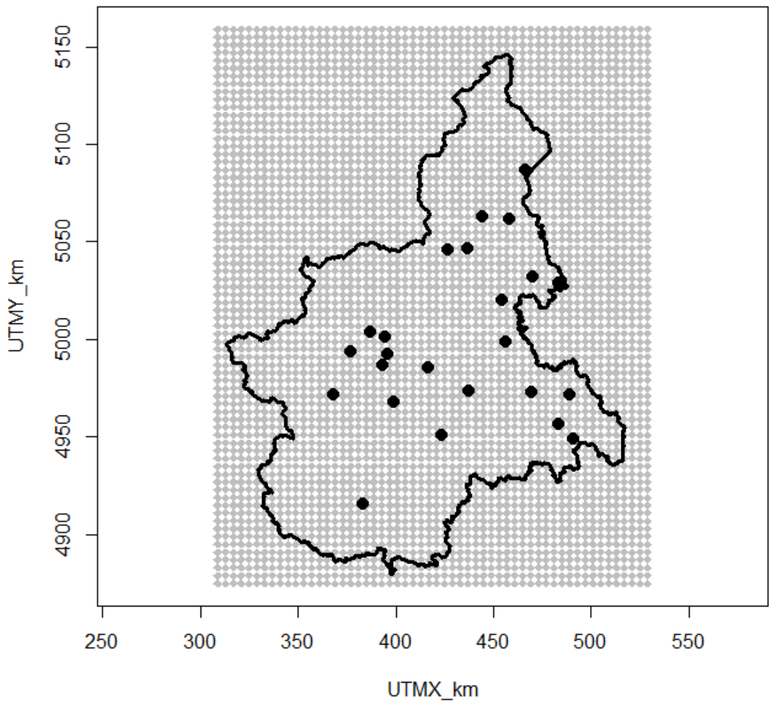
Matérn Spatial Covariance Function

$$\mathrm{Cov}(\xi_i, \xi_j) = \mathrm{Cov}(Z(\xi_i), Z(\xi_j)) = \frac{1}{\Gamma(\lambda)2^{\lambda-1}} (\kappa\|s_i - s_j\|)^{\lambda} K_\lambda(\kappa\|s_i - s_j\|)$$

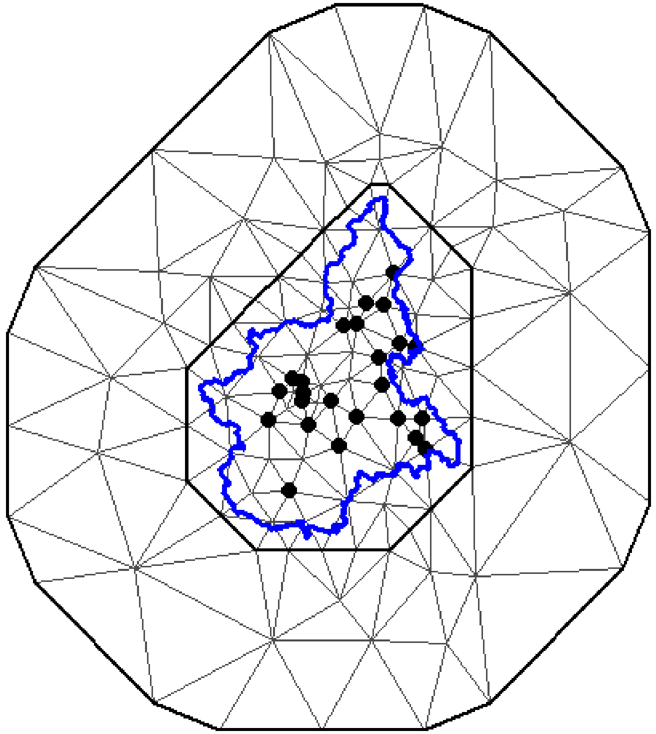
EWHA,
THE FUTURE
WE CREATE

Spatial-temporal Model

Figure 7.8



Mesh Projection



R 코드

obs의 x,y축 좌표

```
Piemonte_mesh = inla.mesh.2d(loc=cbind(coordinates$UTMX,coordinates$UTMY),
                             loc.domain=borders, offset=c(10, 140), max.edge=c(50, 1000))
plot(Piemonte_mesh,asp=1,main="")
lines(borders, lwd=3, col='blue')
points(coordinates$UTMX, coordinates$UTMY, pch=20, cex=2)
```

Mesh 생성

내부, 외부의 비율

내부, 외부에서의
삼각형 변의 최대 길이

EWHA,
THE FUTURE
WE CREATE

Spatial-temporal Model

우리가 알고 싶은 새로운 점(newX,newY)의 좌표

R 코드

Matern 공분산 함수 설정

```
Piemonte_spde = inla.spde2.matern(mesh=Piemonte_mesh, alpha=2)
A_est = inla.spde.make.A(mesh = Piemonte_mesh,
                    loc = coordinates.allyear,        ##obs의 x,y축 좌표
                    group = Piemonte_data$time,       ##같은 날짜 = 같은 group
                    n.group=n_days)                   ##총 182개의 group 생성
s_index = inla.spde.make.index(name = "spatial.field",
                    n.spde = Piemonte_spde$n.spde,    ##mesh 꼭짓점 개수
                    n.group = n_days)
Stack_est <- inla.stack(data = list(logPM10 = Piemonte_data$logPM10),
                    A = list(A_est, 1),               ##Projection Matrix A
                    effects = list(c(s_index, list(Intercept=1)),
                              list(Piemonte_data[,3:10])), tag="est")
```

Projection A matrix 생성

Index 생성

$$\eta_{it} = b_0 + \sum_{m=1}^{M} \beta_m x_{mi} + \omega_{it}$$

```
A_pred = inla.spde.make.A(mesh = Piemonte_mesh,
                    loc = as.matrix(Piemonte_grid),
                    group = i_day,       ##122(꼭짓점 개수)
                    n.group = n_days)
stack_pred = inla.stack(data = list(logPM10 = NA),
                    A = list(A_pred,1),
                    effects = list(c(s_index, list(Intercept=1)),
                              list(covariate_matrix_std)), tag="pred")
                    stack = inla.stack(stack_est, stack_pred)
```

```
formula = logPM10 ~ -1 + Intercept + A + UTMX + UTMY + WS + TEMP
+ HMIX + PREC + EMI +
 f(spatial.field, model=Piemonte_spde,
  group=spatial.field.group, control.group=list(model="ar1"))
output = inla(formula,
        data = inla.stack.data(stack, spde = Piemonte_spde),
        family="gaussian",
        control.predictor=list(A=inla.stack.A(stack), compute=TRUE))
```

$\omega_{it} = a\omega_{i(t-1)} + \xi_{it}$

rep(1:182, each=122)

Model Fitting

Spatial-temporal Model Results

Table 7.2 Posterior estimates (mean, standard deviation (SD), and quantiles) for the Piemonte air pollution model.

Parameter	Mean	SD	2.5%	50%	97.5%
b_0	3.696	0.457	2.784	3.698	4.596
β_1 (A)	−0.209	0.052	−0.313	−0.209	−0.107
β_2 (UTMX)	−0.173	0.170	−0.512	−0.172	0.161
β_3 (UTMY)	−0.179	0.155	−0.487	−0.179	0.125
β_4 (WS)	−0.058	0.008	−0.075	−0.058	−0.042
β_5 (TEMP)	−0.121	0.035	−0.190	−0.120	−0.051
β_6 (HMIX)	−0.025	0.013	−0.051	−0.025	0.001
β_7 (PREC)	−0.054	0.009	−0.071	−0.054	−0.037
β_8 (EMI)	0.035	0.015	0.005	0.035	0.064
σ_e^2	0.032	0.001	0.030	0.032	0.035
σ^2	1.309	0.211	0.956	1.286	1.783
r	269.909	17.034	238.500	269.079	305.408
a	0.960	0.007	0.946	0.960	0.972

$$\eta_{it} = b_0 + \sum_{m=1}^{M} \beta_m x_{mi} + \omega_{it}$$

$$\omega_{it} = a\omega_{i(t-1)} + \xi_{it}$$

PM10가 심한 지역은 중부쪽에 몰려 있을 거란 예측 가능.

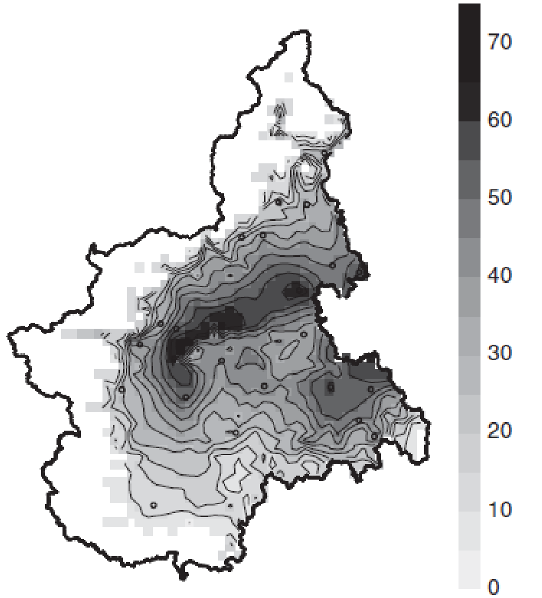


Figure 7.9 Posterior mean of particulate matter concentration for the selected day 30/01/2006. Only locations with an altitude below 1000 m are shown.

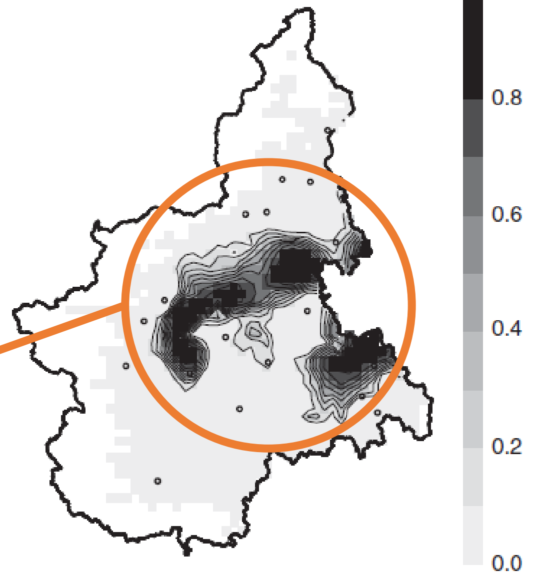


Figure 7.10 Posterior probability of exceeding the 50 μg/m³ limit value for 30/01/2006. Only locations with an altitude below 1000 m are shown.

Spatio-temporal modeling particulate matter concentration

Ⅱ. Change of Support Problem (COSP)

Areal 와 Geostatistical으로 구성된 데이터 (ex. 역학 연구)에서의 공간 불일치를 관리하는 방법론.

Data Explanation

- Geostatistical Data + Areal Data
- 총 12개의 Health District(보건구)가 있으며, 24개의 관측소들이 해당 보건구에 위치함.
- 우리가 예측하고 싶은 new observation (4032개의 좌표) 중 2443개는 보건구 밖에 위치 → 제거

Figure 7.11 Map of the 12 health districts in Piemonte. Black points denote the PM_{10} monitoring stations, while gray diamonds show the 4032 points of the regular grid.

EWHA, THE FUTURE WE CREATE

Change of Support Problem (COSP)

Linear Prediction

$$PM_{B_i t}^{\text{Area}} = \sum_{j \in B_i} \eta_{jt}^{\text{pred}} K_{ij}$$

- S_j : grid location $(\mathrm{j} = 1, \dots, 24)$
- B_i : area $(\mathrm{i} = 1, \dots, 12)$
- $K_{ij} = \frac{1}{\#\,(s_j \in B_i)}$: weight for the prediction

R코드

```
AL_ind = as.numeric(rownames(match_grid_asl[match_grid_asl$COD=="AL",]))
dim_lp = nrow(inla.stack.A(stack)) +ncol(inla.stack.A(stack))

lc_AL_vec = rep(NA, times=dim_lp)
lc_AL_vec[index_pred][AL_ind] = 1/length(AL_ind)
lc_AL = inla.make.lincomb(Predictor = lc_AL_vec)
```

12개의 구역 별로 해당하는 new obs Index 찾기

K_{ij} : Pred의 weight

선형 조합: $\eta_{jt}^{\text{pred}} K_{ij}$

EWHA, THE FUTURE WE CREATE

Change of Support Problem (COSP)

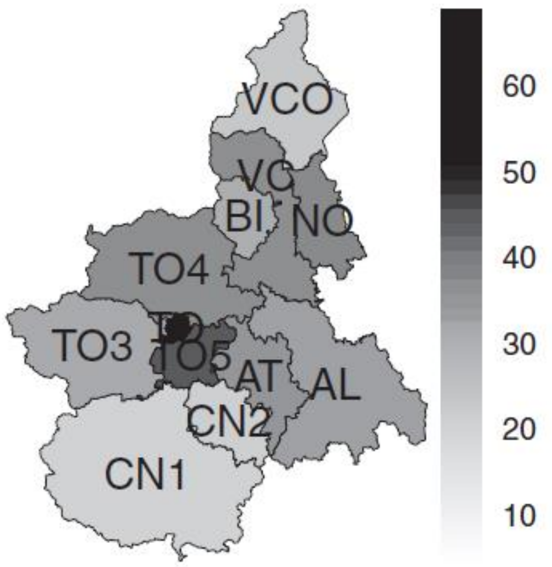


Figure 7.12 Map of the PM_{10} posterior mean at the health district level for 30/01/2006.

: 각 구역별 PM10의 Posterior Mean

→ TO4 보건구의 PM10이 높고, VCO, BI, CN1, CN2는 PM10이 낮을 것으로 예상됨.

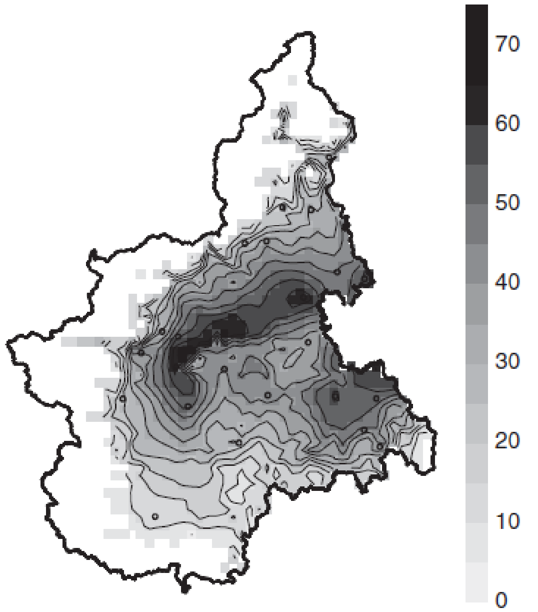


Figure 7.9 Posterior mean of particulate matter concentration for the selected day 30/01/2006. Only locations with an altitude below 1000 m are shown.

EWHA, THE FUTURE WE CREATE