데이터 기반 한강 수질 예측

2018. 06. 07

빅데이터과제 progress seminar

홍 한 움

Datasets

- 수질 일반측정망 (from 물환경정보시스템)
  - **수소이온농도(pH)**, **용존산소량(DO)**, BOD, COD, 부유물질(SS), 총질소(TN), **총인(TP)**, **수온(waterTemp)**, **전기전도도(EC)**, 총대장균군수(Tcol), 용존총질소(DTN), **암모니아성질소(NH3-N)**, **질산성질소(NO3-N)**, 용존총인(DTP), 인산염인(Phosphate), **클로로필-a(Chl-a)**, 분원성대장균군수(fecalCol)
- 기상자료(from 기상자료개방포털)
  - 강우량, 습도, 해면기압(한국정보화진흥원의 낙동강 Chl-a 예측모형에서 주요 변수로 판단)

전처리; 부영양화지수

$TSI_{KO}(COD) = 5.8 + 64.4 \log(COD\ mg/L)$

$TSI_{KO}(CHL) = 12.2 + 38.6 \log(Chl\text{-}a\ mg/m^3)$

$TSI_{KO}(TP) = 114.6 + 43.3 \log(TP\ mg/L)$

위의 세 가지 TSI_{KO}를 종합할 때에는 외부기원 유기물의 지표인 COD에 50%의 가중치를 주고, 내부생성 유기물에 50%의 가중치를 주어 종합 TSI_{KO}를 계산한다. 내부생성유기물의 지표는 조류의 밀도지표인 Chl-a이며 TP는 조류의 밀도를 좌우하는 지표이므로 이 두 가지에 각각 25%의 가중치를 주어 다음과 같이 계산하면 된다.

$종합\ TSI_{KO} = 0.5\ TSI_{KO}(COD) + 0.25\ TSI_{KO}(CHL) + 0.25\ TSI_{KO}(TP)$

출처: 환경부(2006), 물환경종합평가방법 개발 조사연구(III) 최종보고서
– 부영양화조사 및 평가체계 연구

전처리

2010-01-07 부터 2017-12-30 까지 전체 있어야할 자료 수 : 417개

TOC 를 설명변수로 활용하는 것을 포기하고 2010-1월부터 분석 vs TOC, Penol 을 설명변수로 활용하고 2011-9-10부터 분석 (즉, 자료수 87개 vs 설명변수 1개)

위치별 자료수

	location	freq
5	노량진	403
105	섬강4-1	403
47	경안천5	400
192	팔당댐	398
116	안성천3	393
39	강천	390
152	이포	390
1	가양	389
48	경안천5A	384
38	강상	381
103	삼봉리	381
158	임진강4	366
44	경안천3A	349
128	여주1	343
129	여주2	339
15	안양천4	334
23	중랑천1A	334
37	가평천3	334
60	공릉천3	334

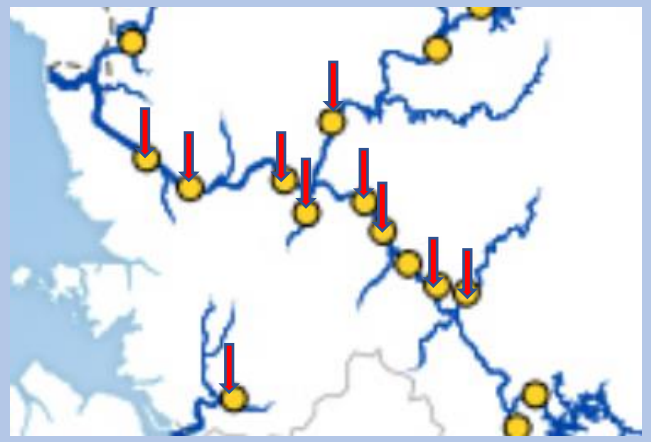
수질측정주요지점 물환경정보시스템

메인페이지
(http://water.nier.go.kr/main/mainContent.do)

가양, 노량진, 팔당댐, 경안천5, 삼봉리, 강상, 이포, 강천, 섬강4-1, 안성천3

전처리

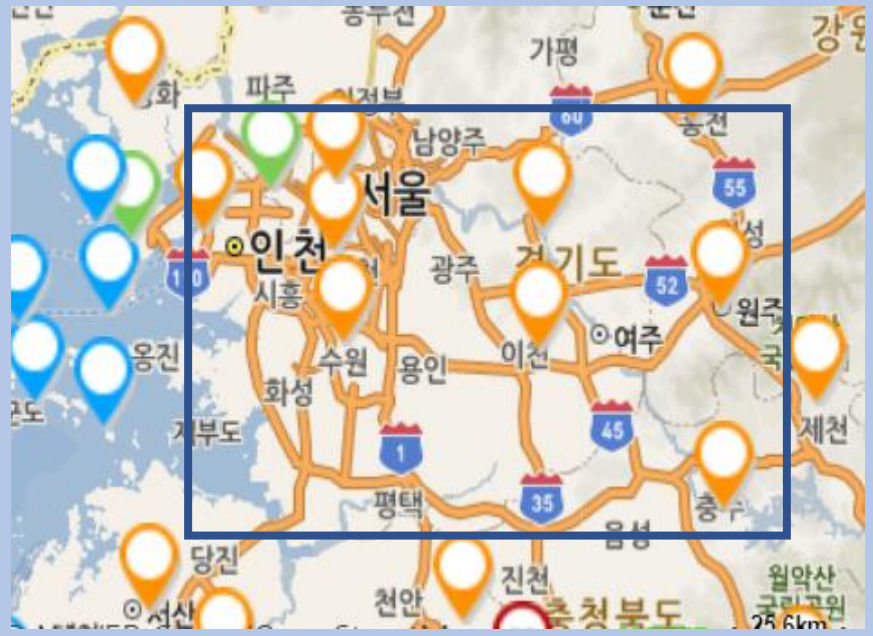
수질자료 관측지점



가양, 노량진, 팔당댐, 경안천5, 삼봉리, 강상, 이포, 강천, 섬강4-1, 안성천3

location	longitude	latitude
경안천5	127.304817	37.436144
팔당댐	127.285051	37.527818
삼봉리	127.326546	37.597443
강상	127.488532	37.474873
이포	127.552017	37.402458
강천	127.676340	37.269077
섬강4-1	127.746065	37.242565
안성천3	127.088484	36.982368
가양	126.844667	37.566445
노량진	126.966518	37.517459

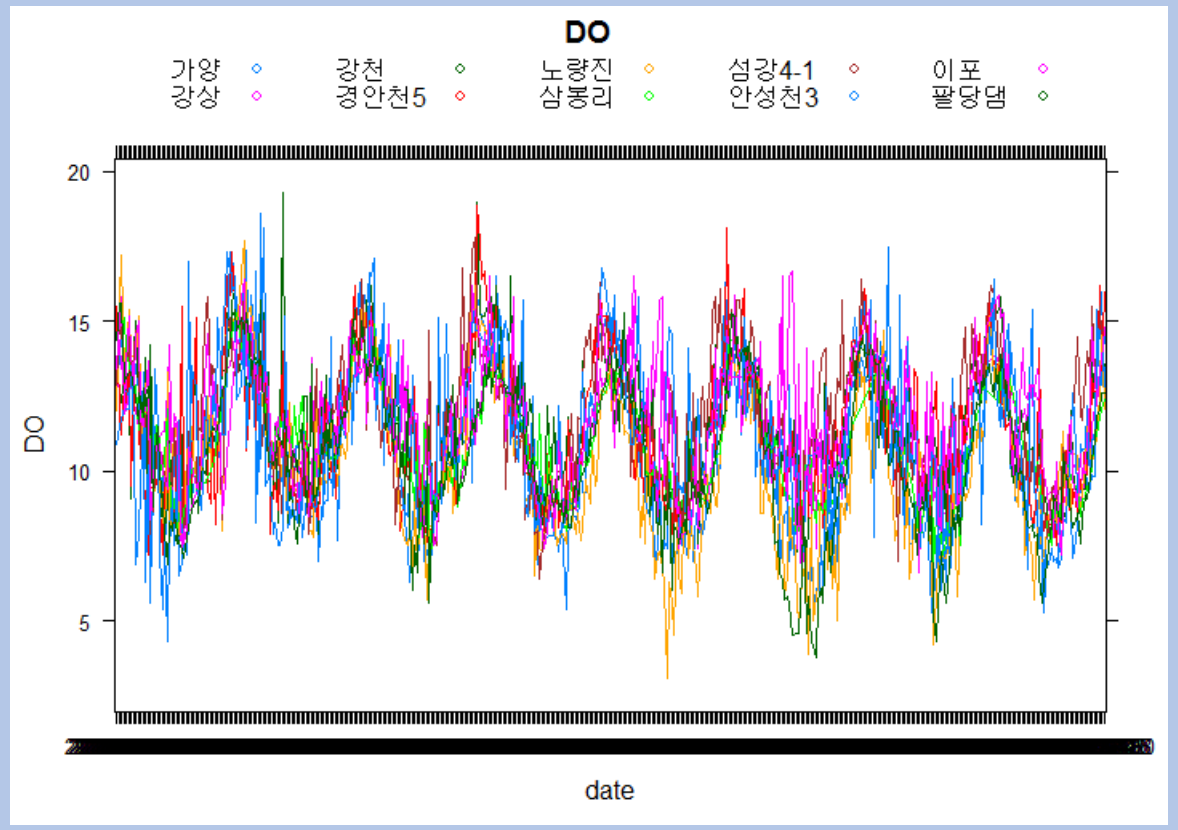
기상자료 관측지점:

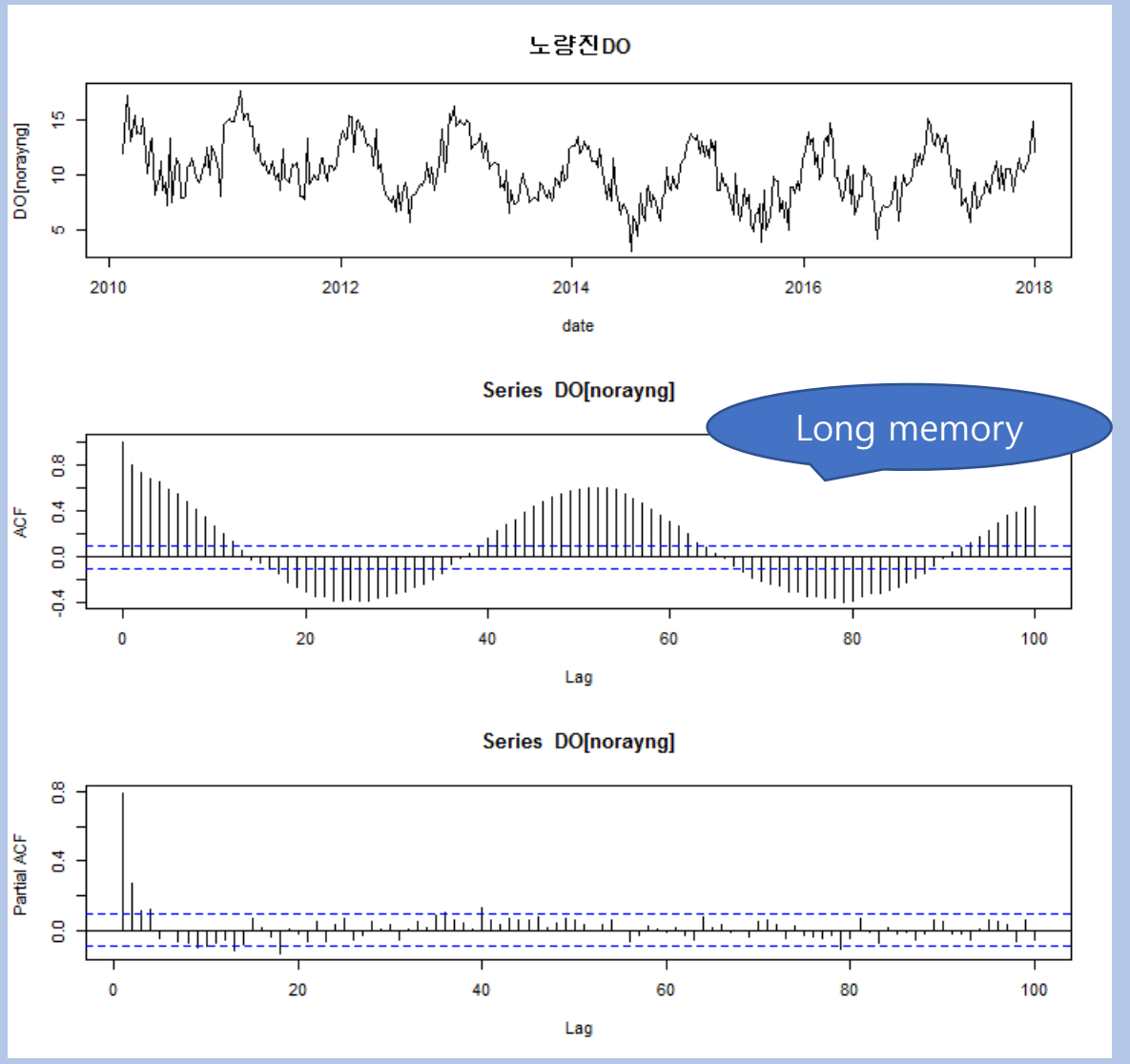


인천, 서울, 관악 수원, 양평, 이천, 원주, 충주, 천안

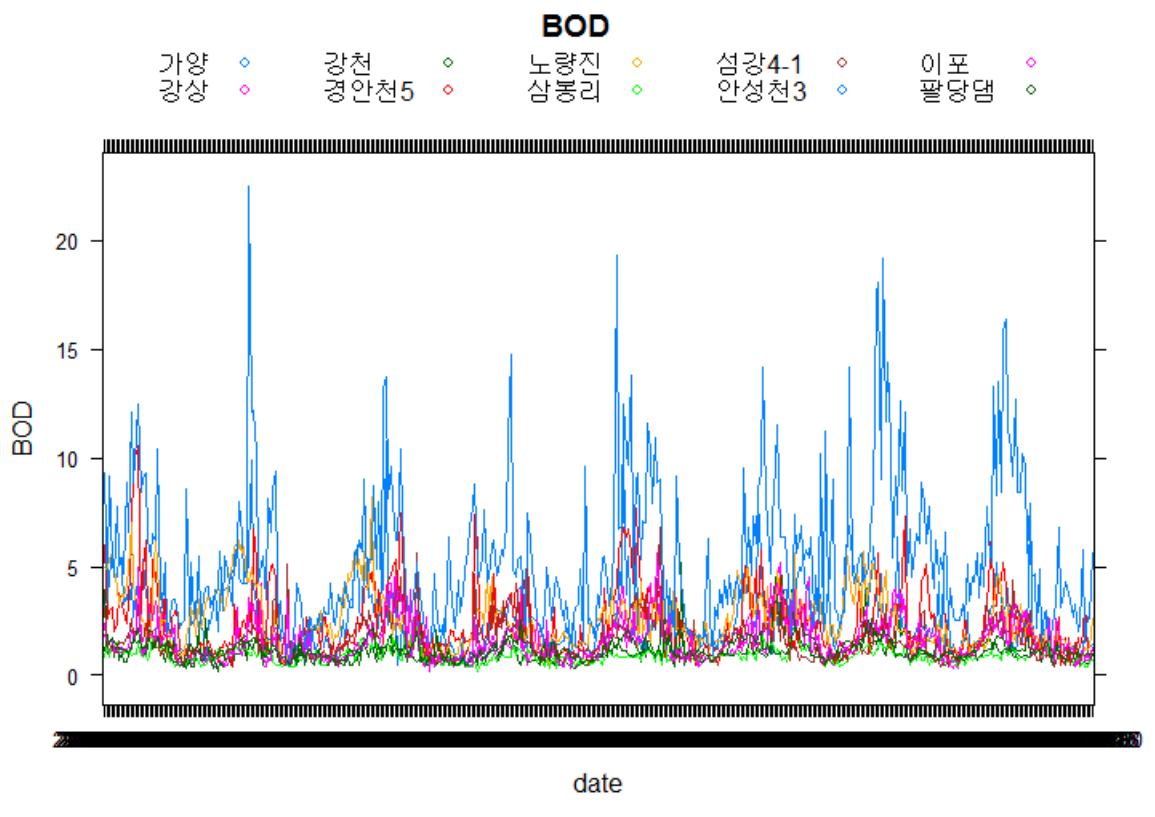
location	지점	longitude	latitude
관악	116	126.964	37.4453
서울	108	126.9658	37.5714
인천	112	126.6249	37.4777
수원	119	126.9853	37.2723
양평	202	127.4945	37.4886
이천	203	127.4842	37.264
원주	114	127.9466	37.3376
충주	127	127.9527	36.9704
제천	221	128.1943	37.1593
천안	232	127.1191	36.7796

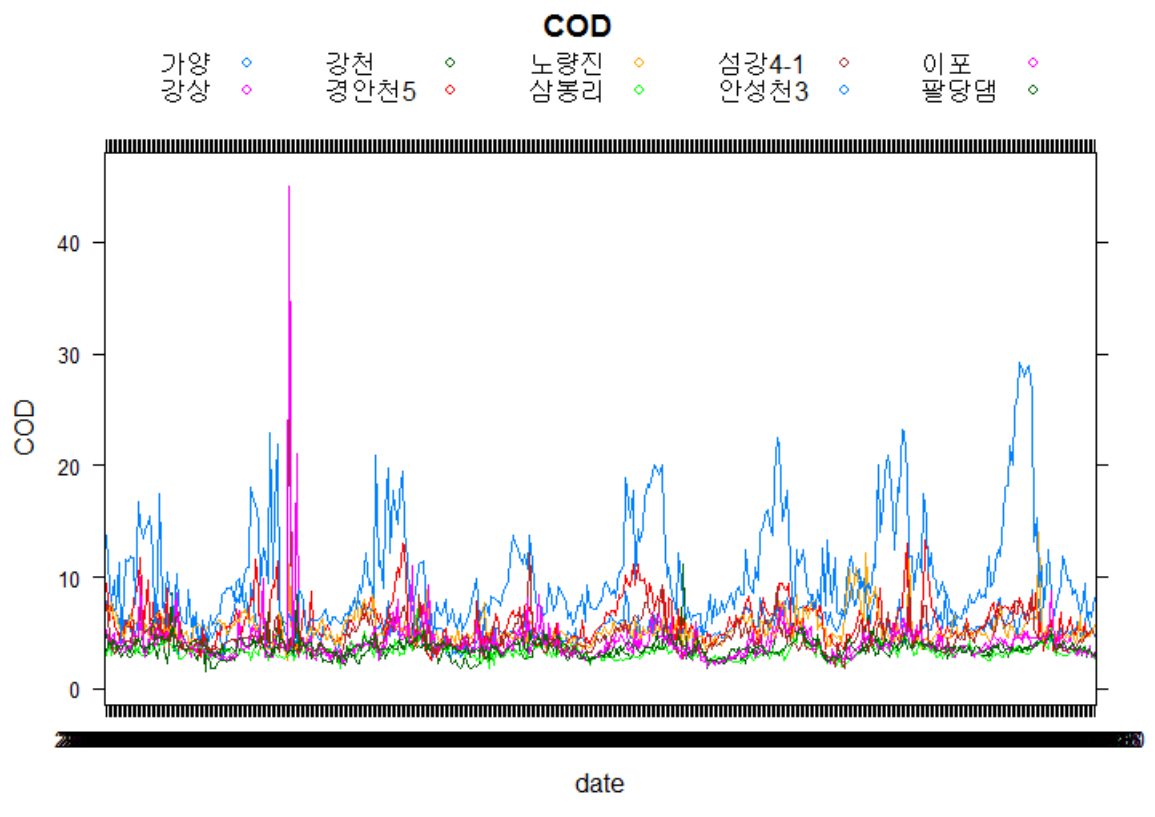
기술통계: 변수 Graph



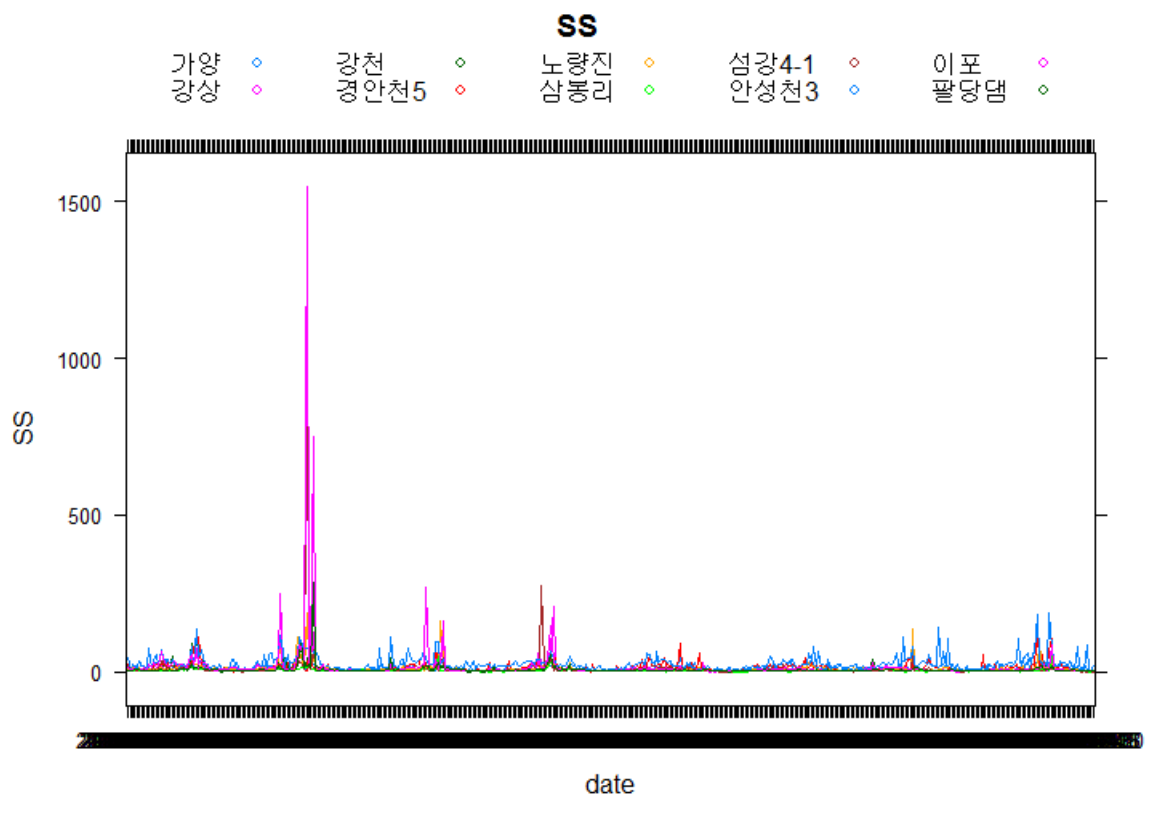


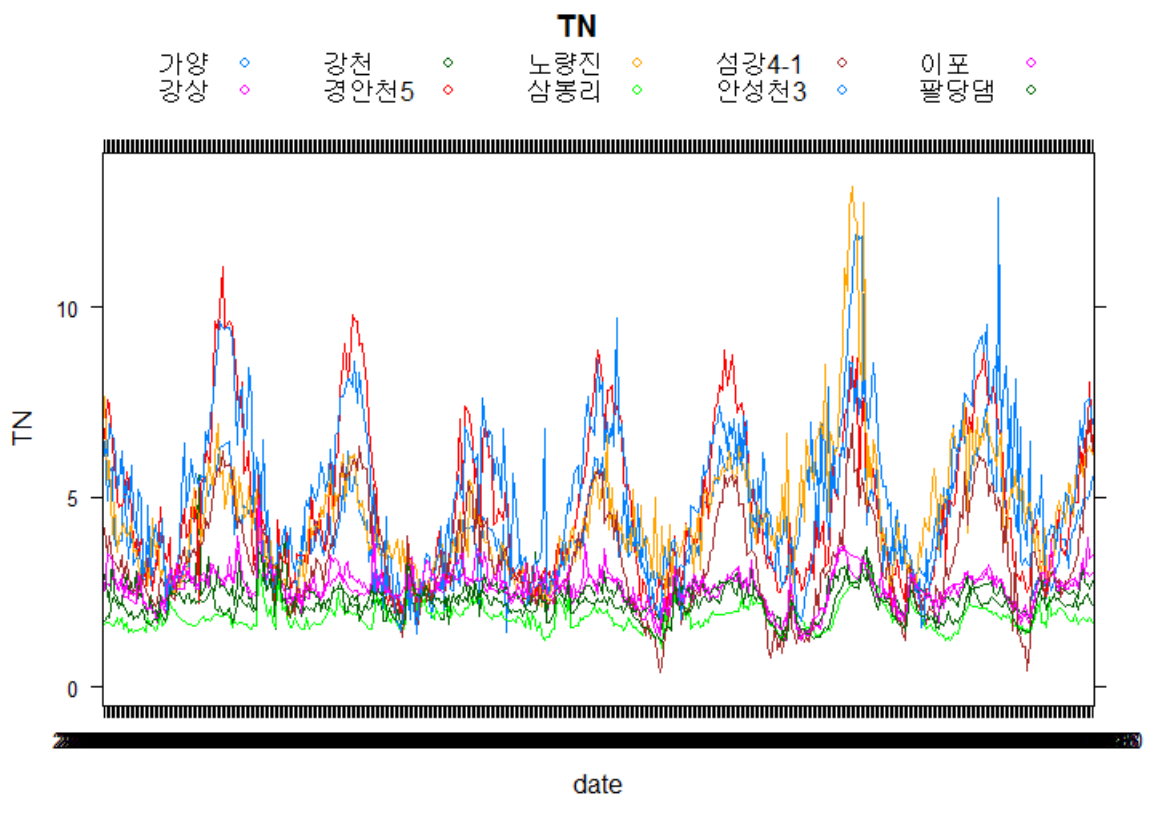
기술통계: 변수 Graph



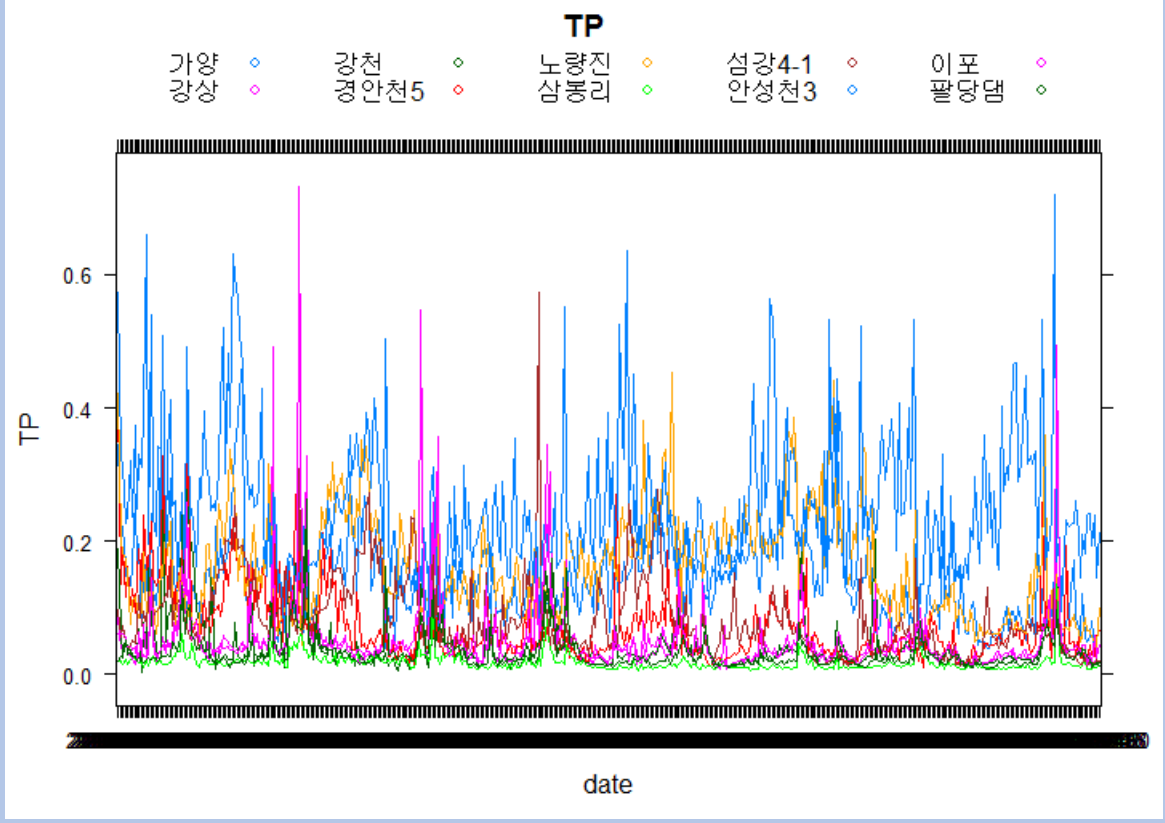


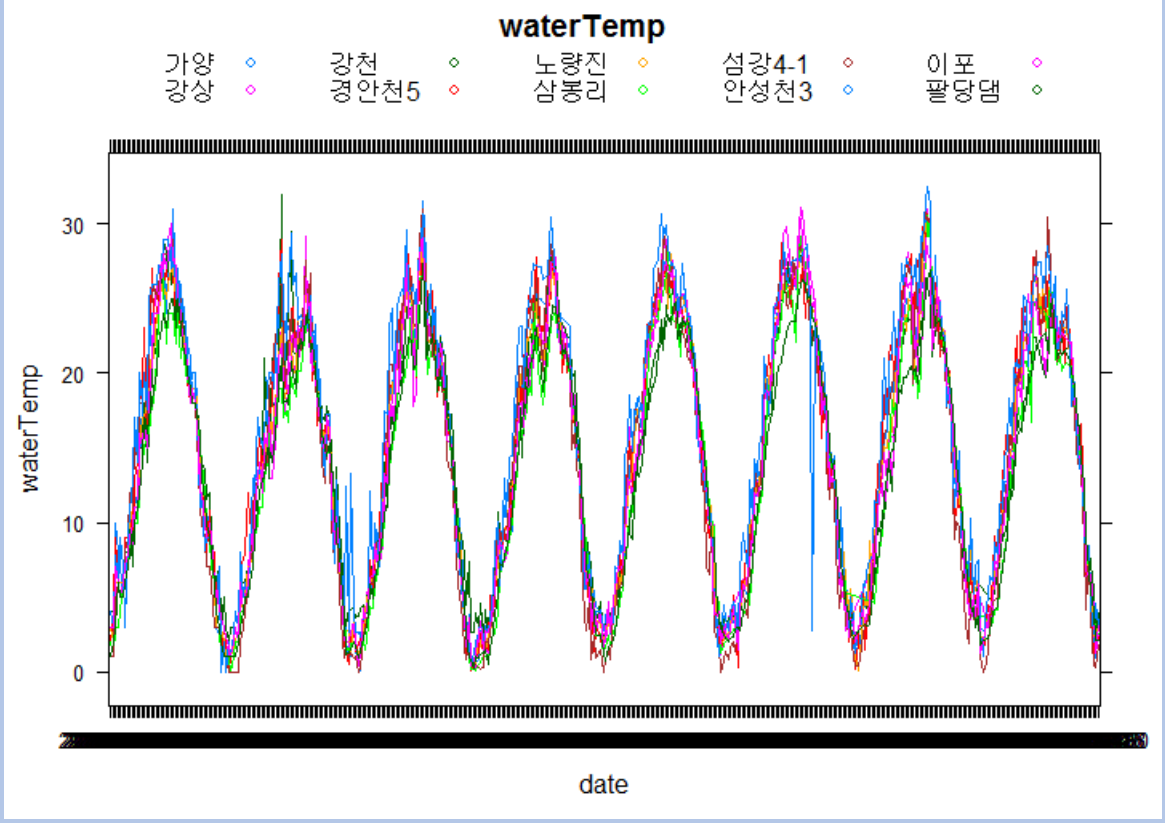
기술통계: 변수 Graph



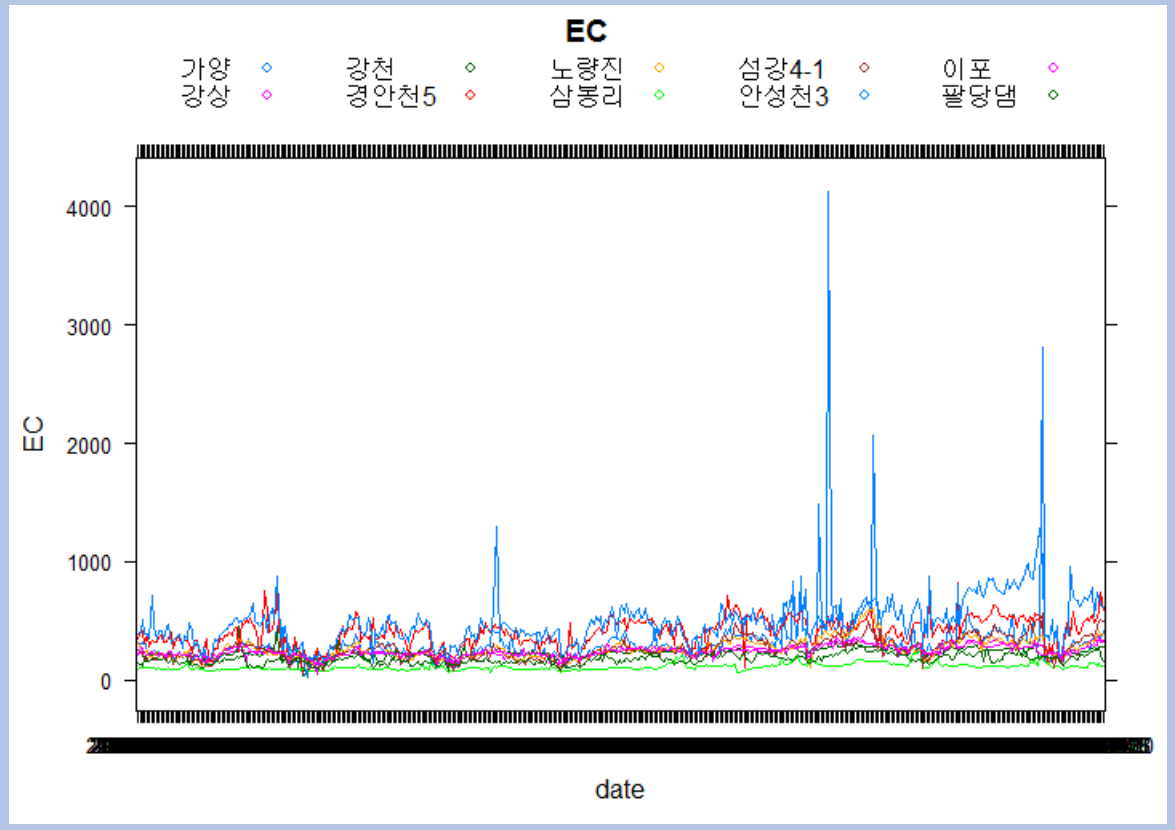


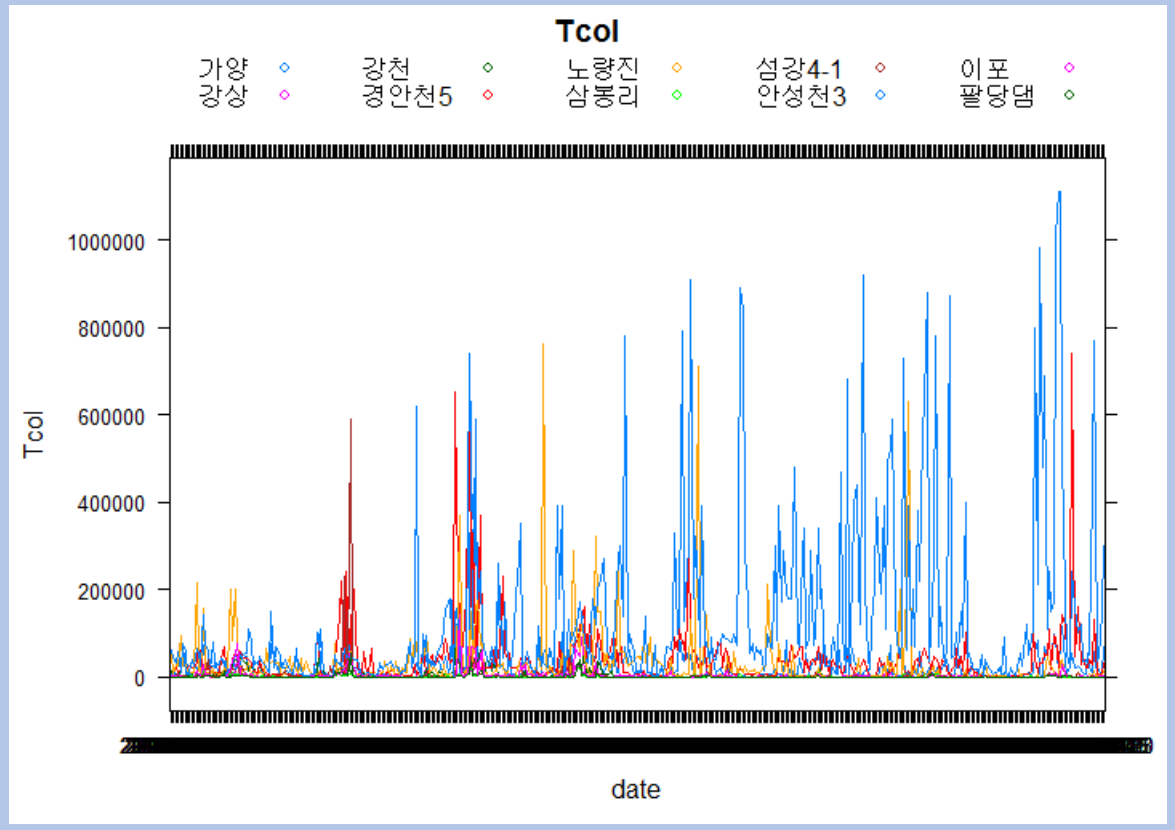
기술통계: 변수 Graph



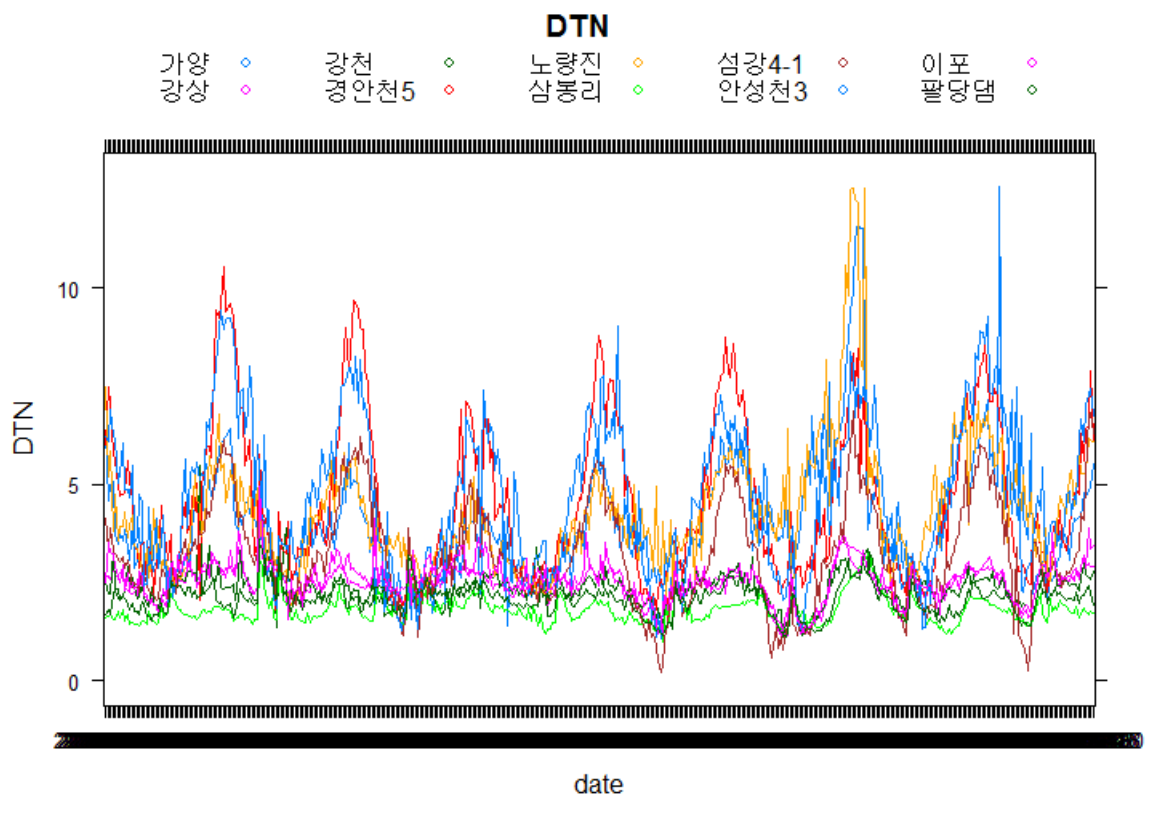


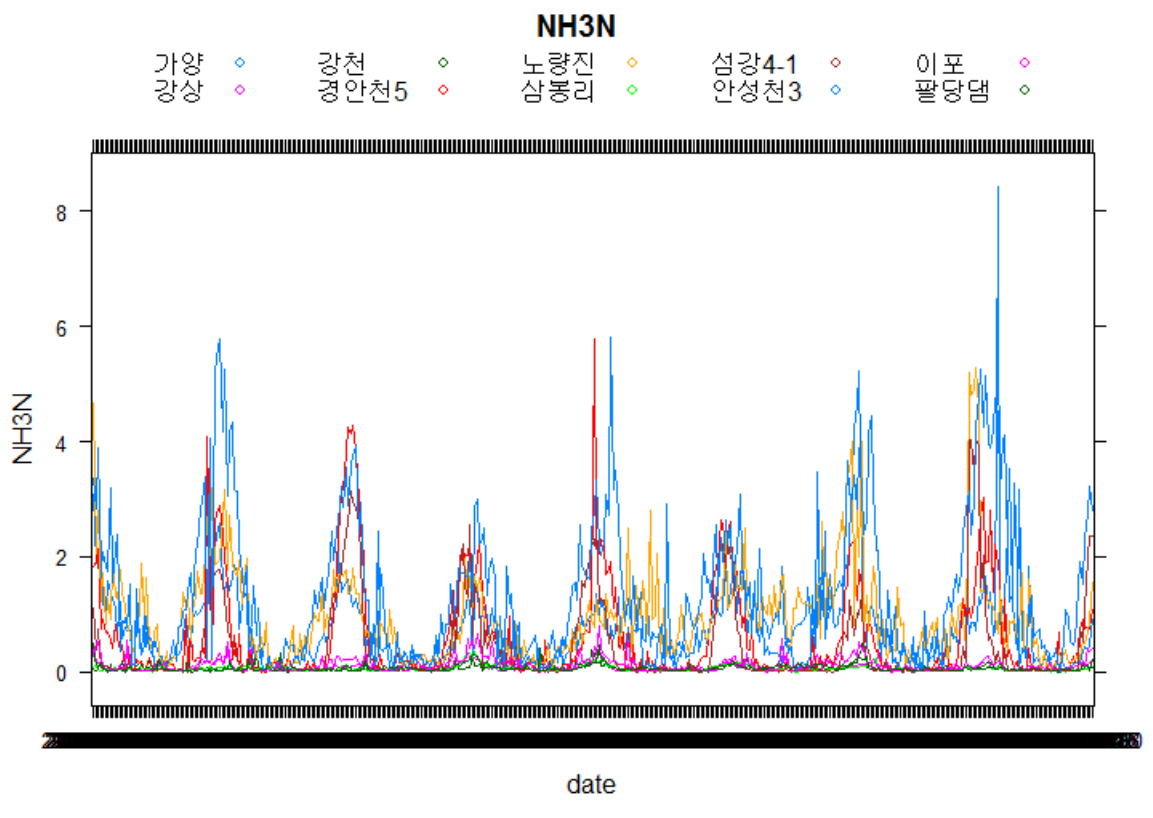
기술통계: 변수 Graph



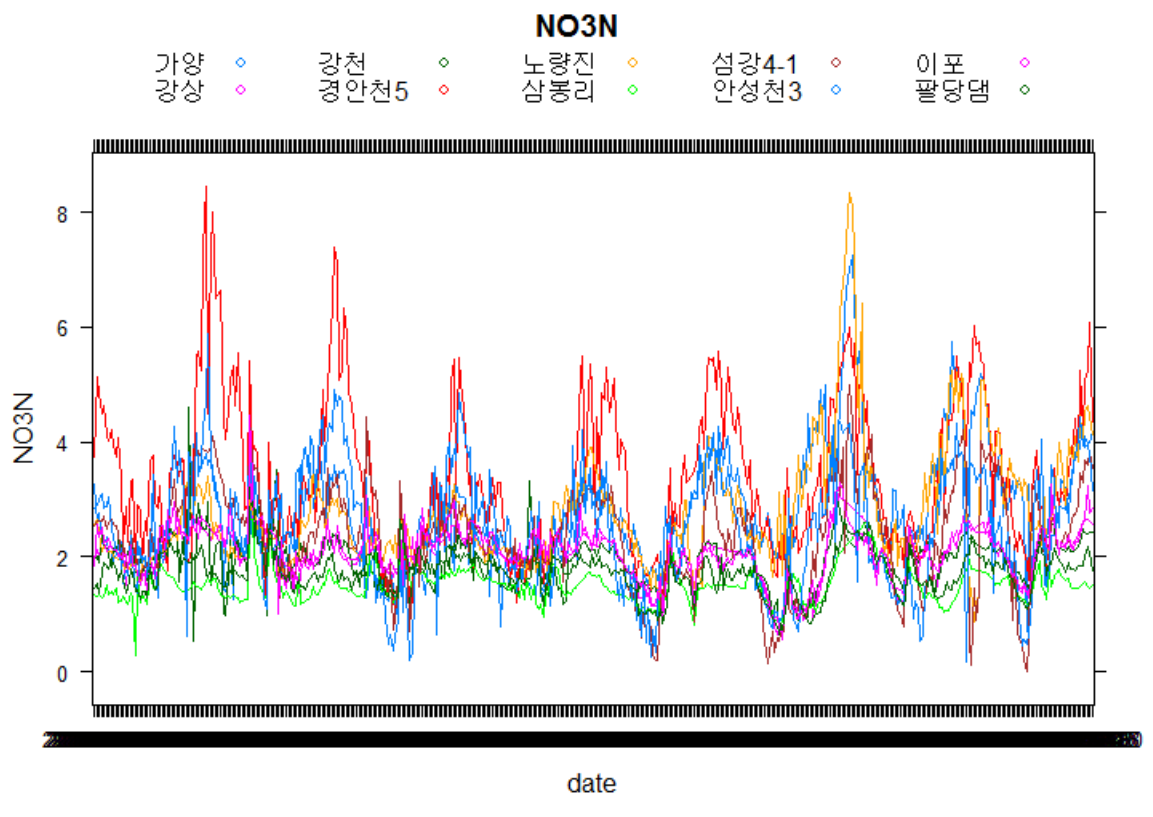


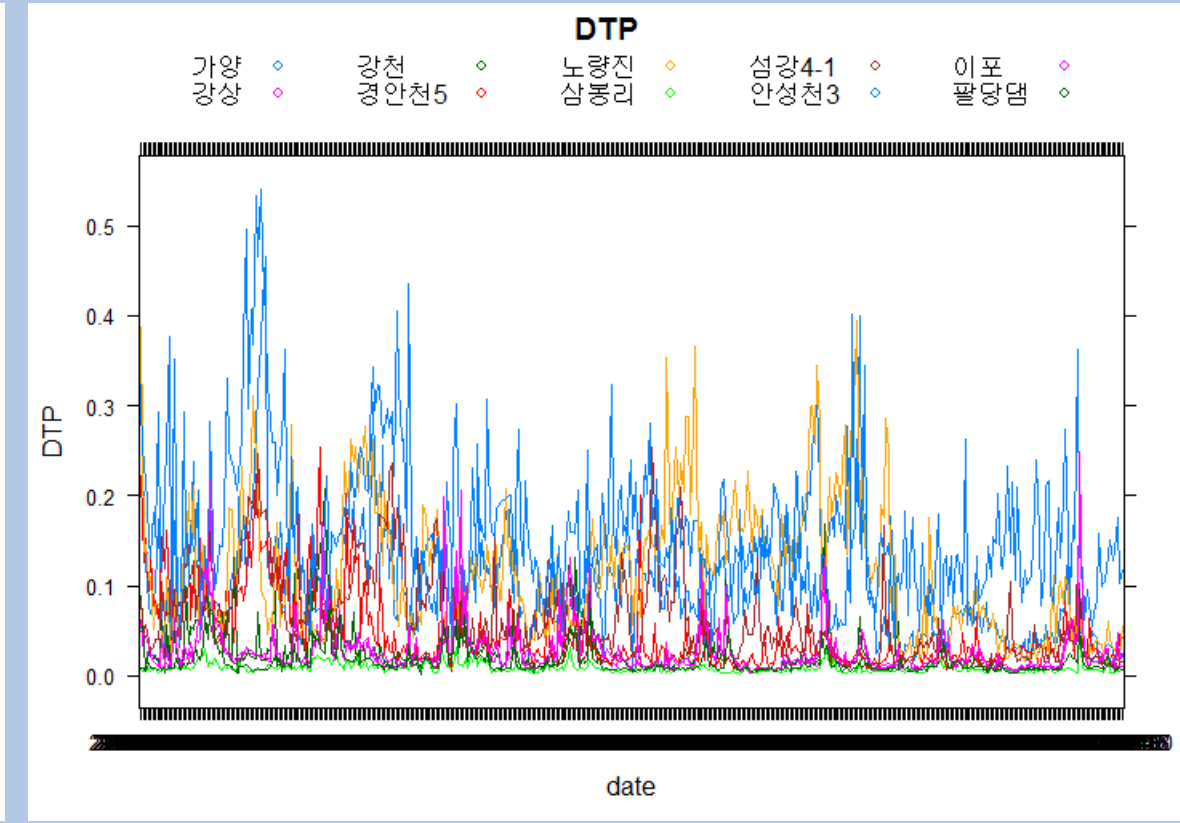
기술통계: 변수 Graph



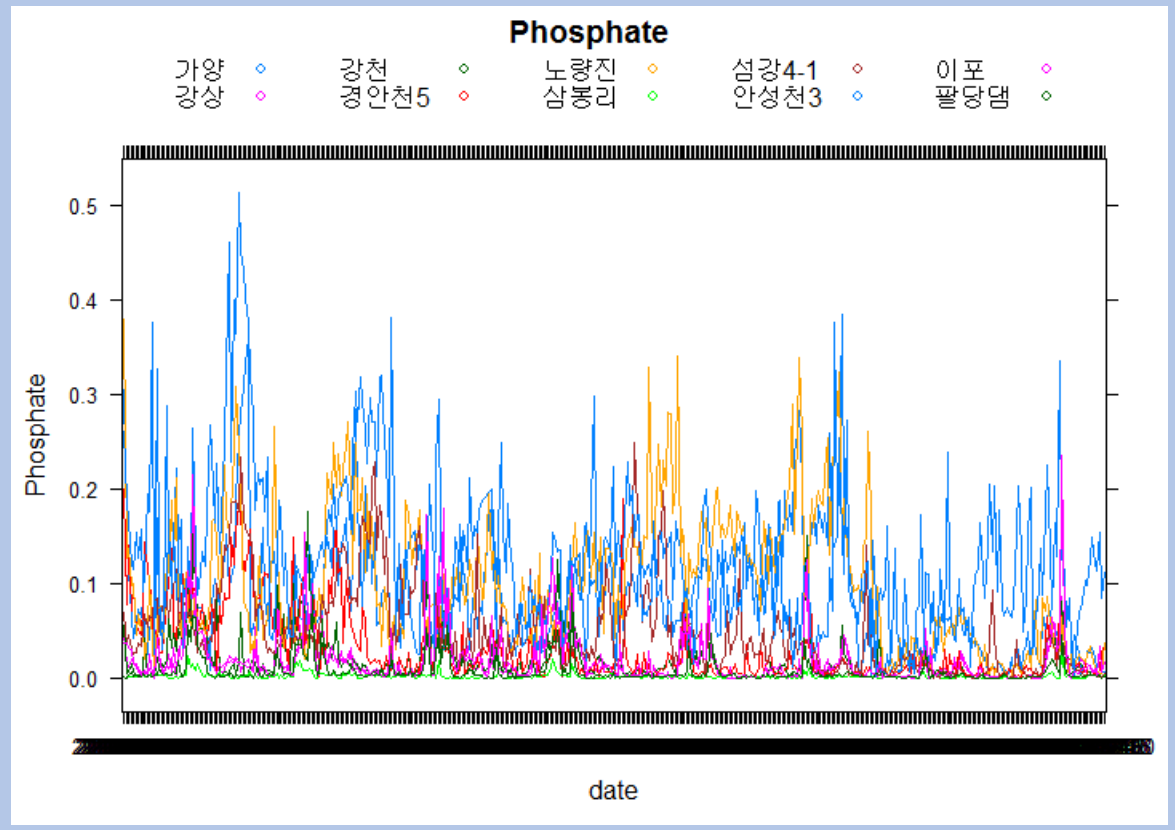


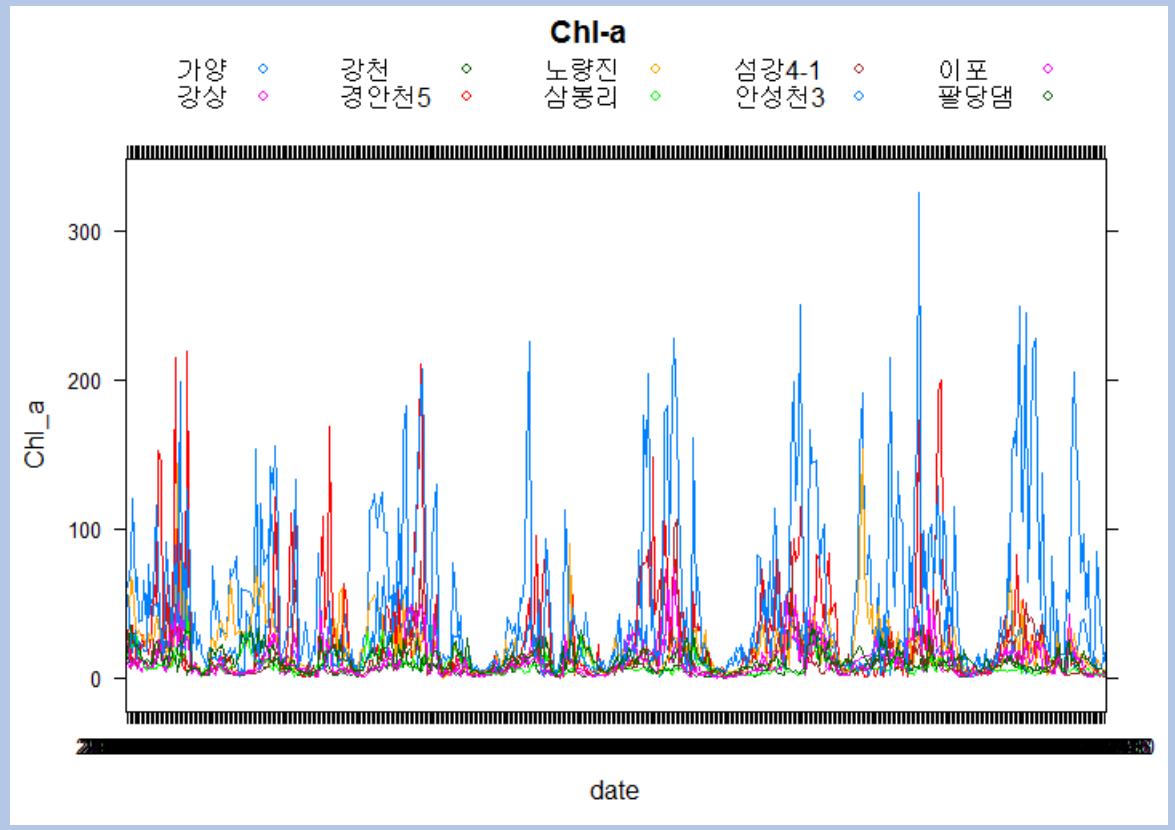
기술통계: 변수 Graph



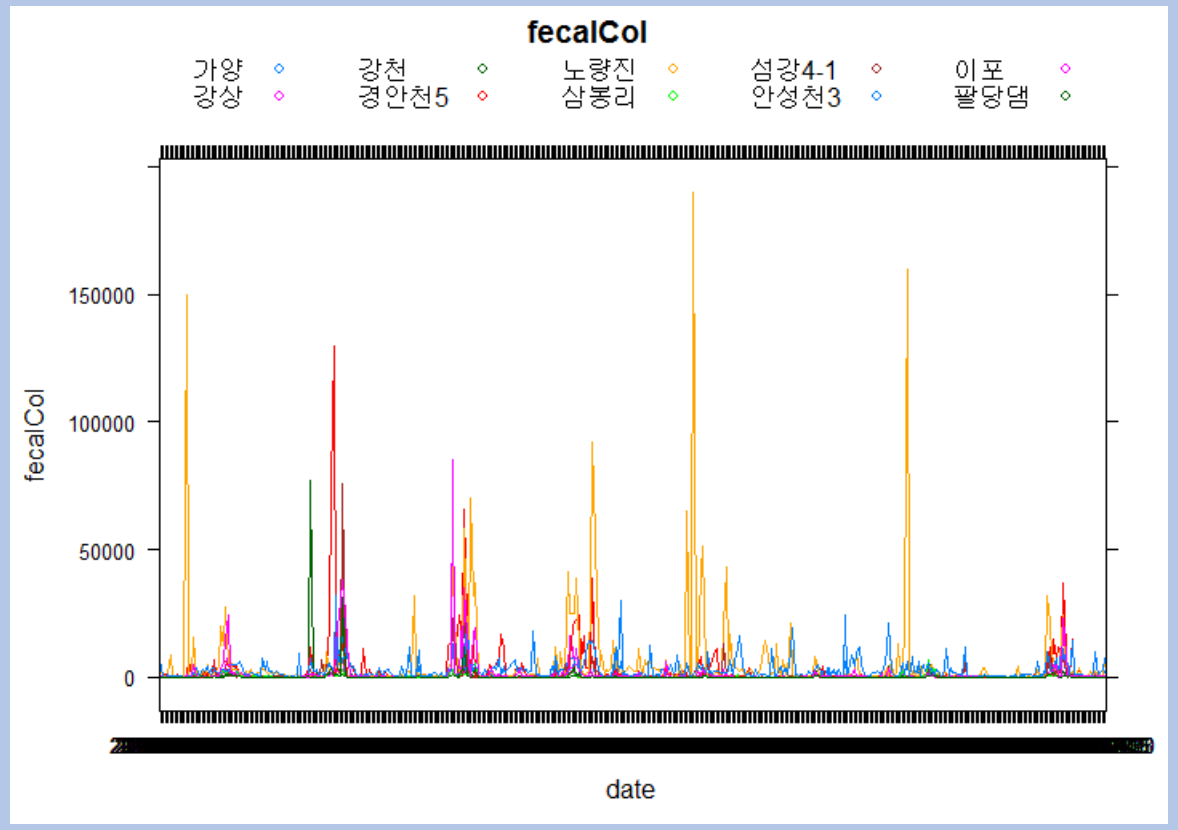


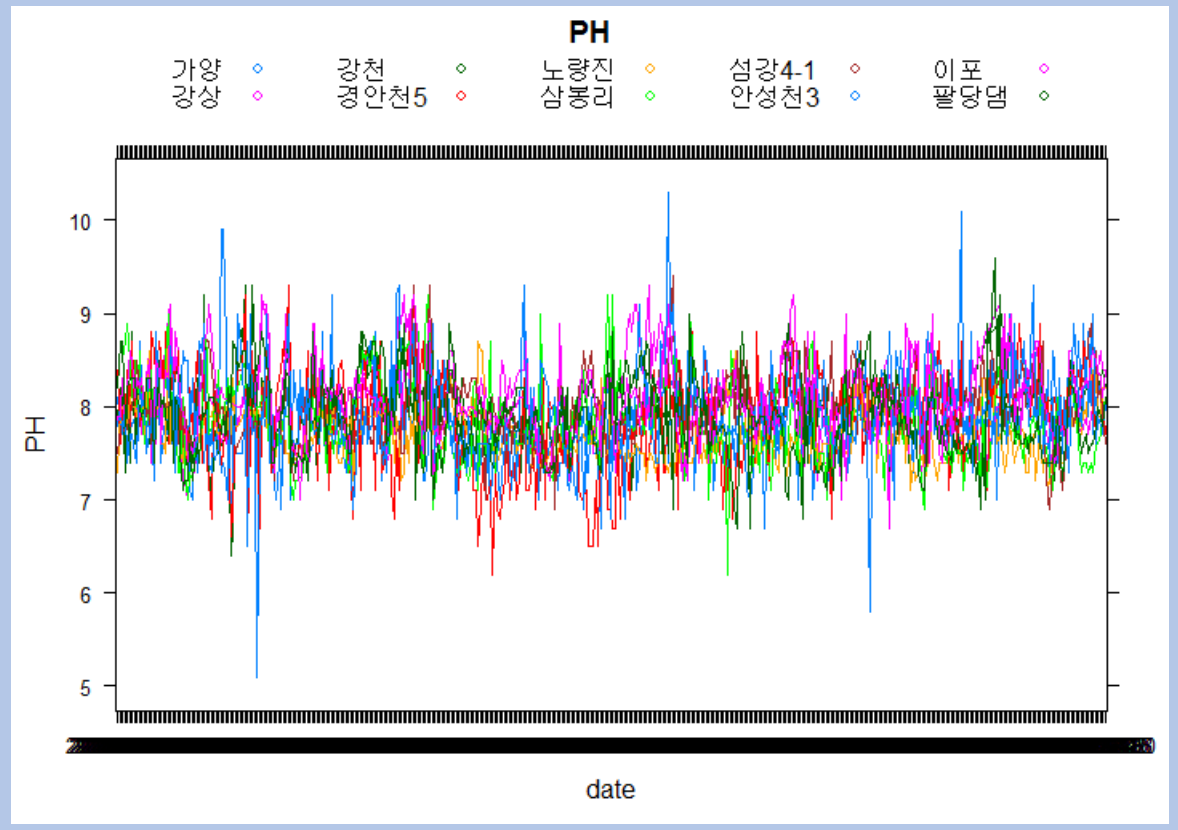
기술통계: 변수 Graph





기술통계: 변수 Graph





Methodology

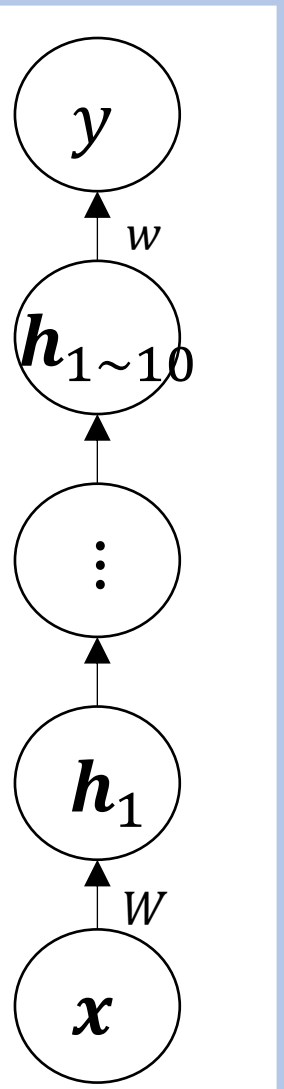
1. ANN
2. RNN, GRU (or LSTM)
3. KNN
4. 시공간자료분석모형
5. AutoEncoder

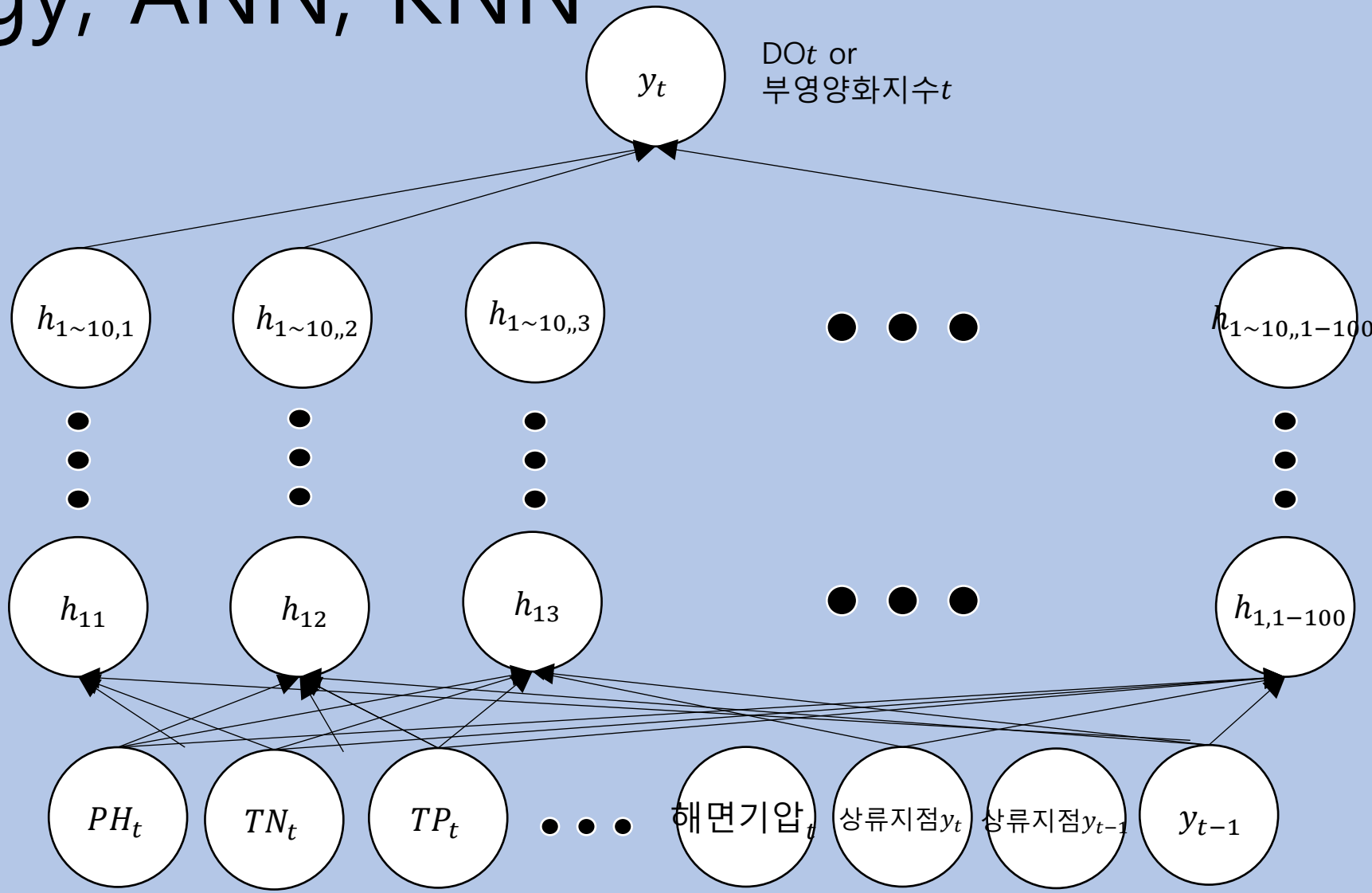
- 분석언어: python tensorflow

Methodology; ANN, KNN

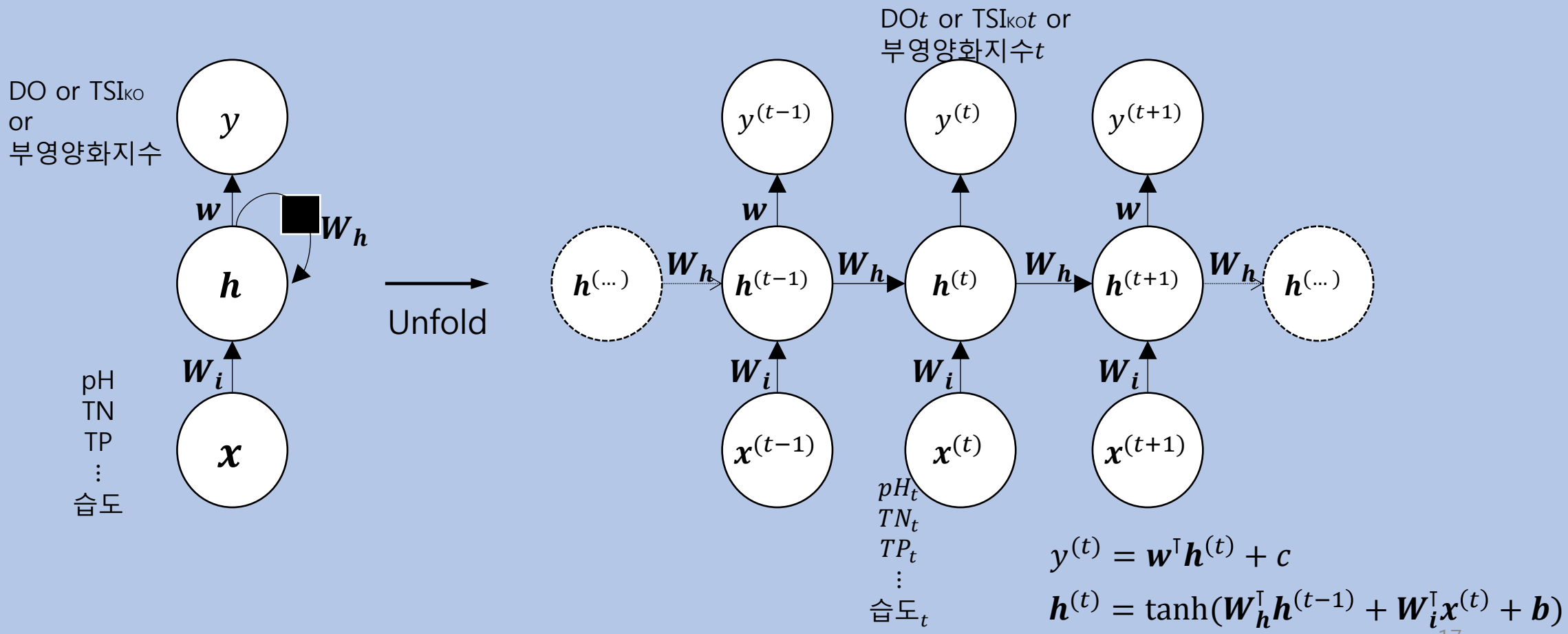
$y = \boldsymbol{w}^{\intercal}\boldsymbol{h} + \boldsymbol{c}$

$\boldsymbol{h} = tanh(\boldsymbol{W}^{\intercal}\boldsymbol{x} + \boldsymbol{b})$
$\boldsymbol{or}, \max\{0, \boldsymbol{W}^{\intercal}\boldsymbol{x} + \boldsymbol{b}\}$





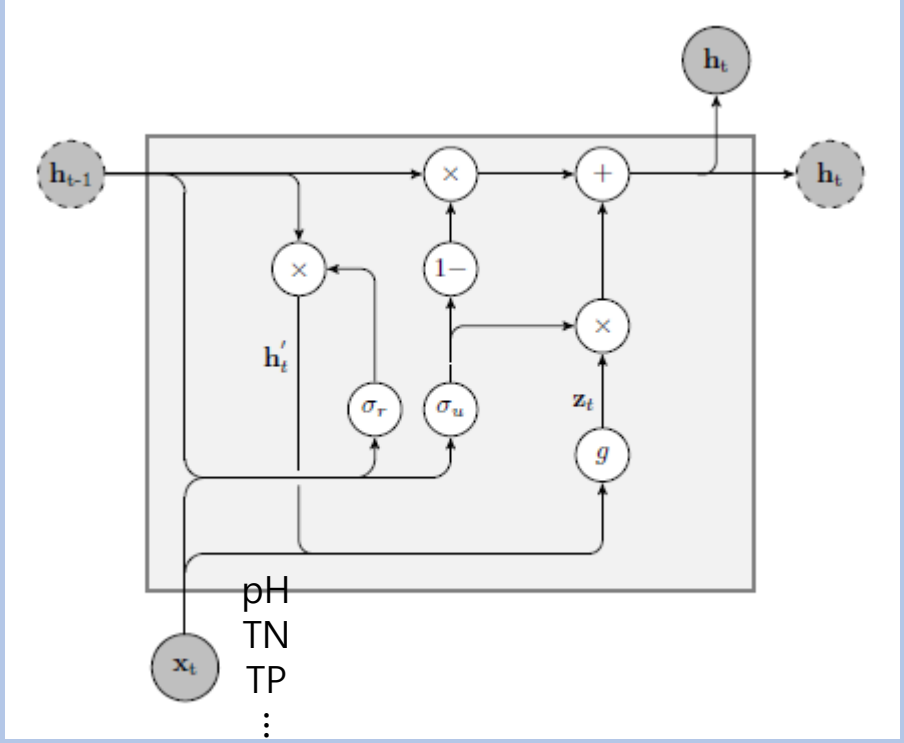
Methodology; RNN



$$y^{(t)} = \boldsymbol{w}^{\intercal}\boldsymbol{h}^{(t)} + c$$

$$\boldsymbol{h}^{(t)} = \tanh(\boldsymbol{W}_{\boldsymbol{h}}^{\intercal}\boldsymbol{h}^{(t-1)} + \boldsymbol{W}_{\boldsymbol{i}}^{\intercal}\boldsymbol{x}^{(t)} + \boldsymbol{b})$$

Methodology; GRU



$$
\begin{aligned}
\text{reset gate}: \quad & \mathbf{r}[t] = \sigma\left(\mathbf{W}_r \mathbf{h}[t-1] + \mathbf{R}_r \mathbf{x}[t] + \mathbf{b}_r\right), \\
\text{current state}: \quad & \mathbf{h}'[t] = \mathbf{h}[t-1] \odot \mathbf{r}[t], \\
\text{candidate state}: \quad & \mathbf{z}[t] = g\left(\mathbf{W}_z \mathbf{h}'[t-1] + \mathbf{R}_z \mathbf{x}[t] + \mathbf{b}_z\right), \\
\text{update gate}: \quad & \mathbf{u}[t] = \sigma\left(\mathbf{W}_u \mathbf{h}[t-1] + \mathbf{R}_u \mathbf{x}[t] + \mathbf{b}_u\right), \\
\text{new state}: \quad & \mathbf{h}[t] = (1 - \mathbf{u}[t]) \odot \mathbf{h}[t-1] + \mathbf{u}[t] \odot \mathbf{z}[t].
\end{aligned}
$$