데이터 기반 한강 수질 예측

2018. 05. 10

빅데이터과제 progress seminar

홍 한 움

전처리

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

2011-09-10 부터 2017-12-30 까지 전체 있어야할 자료 수 : 330개

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

전체 자료 대상 위치별 자료수

	location	freq
69	섬강4-1	403
4	노량진	402
129	팔당댐	398
31	강상	380
68	삼봉리	380
101	이포	368
37	경안천5	335
1	가양	327
76	안성천3	327
32	강천	325
105	임진강4	314
53	대신	294
83	여주1	193
84	여주2	191
148	아라천1	101
11	안양천4	96
30	가평천3	96
43	공릉천3	96
73	신천3	96

DO 자료 대상 위치별 자료수

	location	freq	
5	노량진	403	
105	섬강4-1	403	
47	경안천5	400	<- TOC, Penol 에 missing
192	팔당댐	398	
116	안성천3	393	<- TOC, Penol 에 missing
39	강천	390	<- TOC, Penol 에 missing
152	이포	390	<- Penol 에 missing
1	가양	389	<- TOC, Penol 에 missing
48	경안천5A	384	<- Penol, Tcol, DTN, NH3N, DTP, Phosph, Chl-a, fetalCo 관측값 없음
38	강상	381	
103	삼봉리	381	
158	임진강4	366	<- TOC, Penol 에 missing
44	경안천3A	349	
128	여주1	343	
129	여주2	339	
15	안양천4	334	
23	중랑천1A	334	
37	가평천3	334	
60	공릉천3	334	

TOC 에 missing 이 있는 자료들은 2011년 9월부터 제대로 관측되기 시작됨

전처리

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

선행연구

- 한국정보화진흥원
  - 낙동강 유역 녹조 발생 예측
  - SVM, Random forest, RNN, 선형회귀모형
- Rankovic et al. (2012)
  - 딥러닝 기반 Gruža 저수지 용존산소량 예측
- Xu et al. (2012)
  - 시공간분석 기반 Zhangweinan 강 용존산소량 분석

Datasets

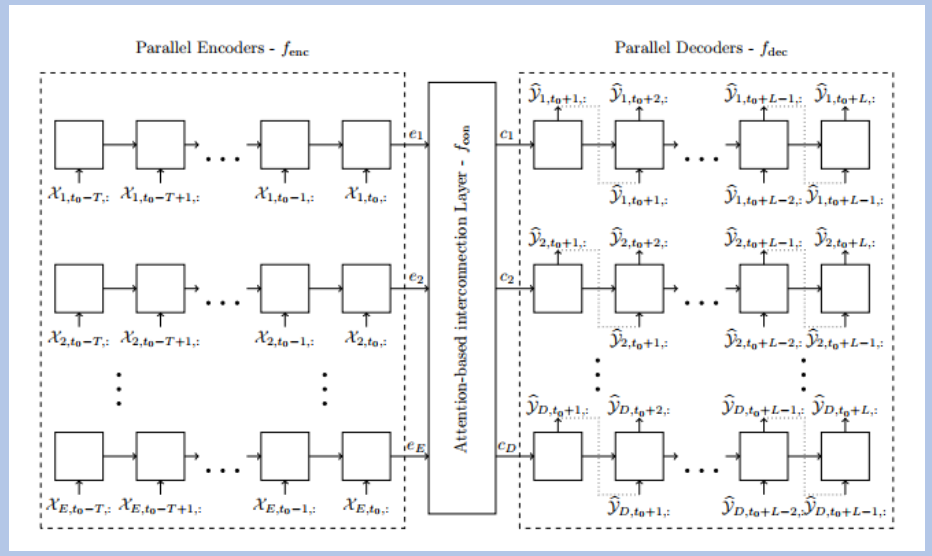
- 수질 일반측정망 (from 물환경정보시스템)
  - **수소이온농도(pH)**, **용존산소량(DO)**, BOD, COD, 부유물질, 총질소(TN), **총인(TP)**, TOC, **수온**, **전기전도도**, 총대장균군수, 용존총질소, **암모니아성질소(NH3-N)**, **질산성질소(NO3-N)**, 용존총인, 인산염인, **클로로필-a**, 분원성대장균군수
- 기상자료
  - **강우량**, **습도**, 기온
- 2010-2017 수도권 수질 측정지역 주해상도 자료이용

가용자료 개수	측정소 수
400이상	3
350-400	9
300-350	24
200-300	21
151미만	159
합계	216

- 350개 이상의 자료를 사용 가능한 12개 측정소 대상으로 예측 진행
- 나머지 자료는 설명변수로 활용

Methodology

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



- 분석언어: python tensorflow
- 선형회귀모형, VARMA 대비 우수한 예측 결과 예상

예상소요기간

- 자료수집 및 전처리 ; 3월
  - 도메인 분석
  - 결측값 보간
  - 이상치 제거/대체
- 자료분석 ; 4-8월
  - Simple ANN, RNN ; 4-6월
  - GRU, KNN ; 5-6월
  - 시공간자료분석 ; 5-6월 //중간보고
  - AutoEncoder ; 7-8월
- 보고서 작성 ; 6-10월