딥러닝을 활용한 시공간 데이터 예측
Deep Learning for Spatial-Temporal Prediction and Forecast

박태준(2021-24984), 진수정(2022-22742)

딥러닝의 통계적 이해

2022년 겨울

목차

Section 1

개요

1. 개요

- 이 프로젝트를 통해 우리는 시공간 데이터(Spatio-temporal data)에 대한 예측(Prediction, Interpolation)과 미래 값에 대한 예측(Forecast, Outerpolation)을 위한 새로운 DNNs(Deep neural networks) 모델 구조를 제안한다.

1. 개요

시공간 크리깅(Spatio-Temporal Kriging)

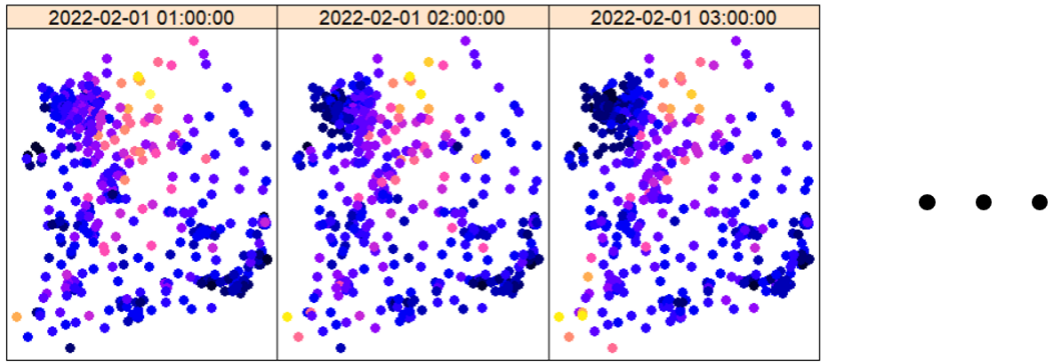


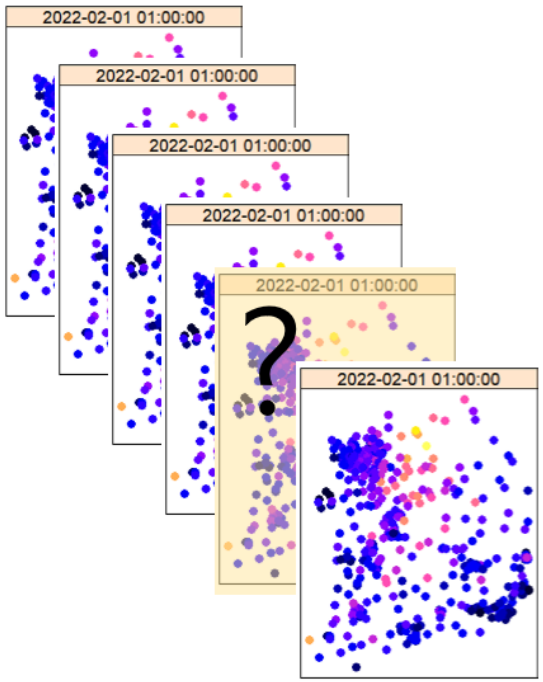
Figure: 시공간 데이터 예시

- 시공간 크리깅이란 시공간적 의존성(Dependence)를 이용한 관측되지 않은 장소와 시간에 대한 값의 예측이다.
- 크리깅 예측은 확률과정의 가우시안(Gaussian) 가정과 시공간 공분산 함수(Covariance function)의 정상성(Stationarity) 가정 하에서 도출된다.
- 관측 데이터의 수가 매우 많으면 계산이 힘들어진다.

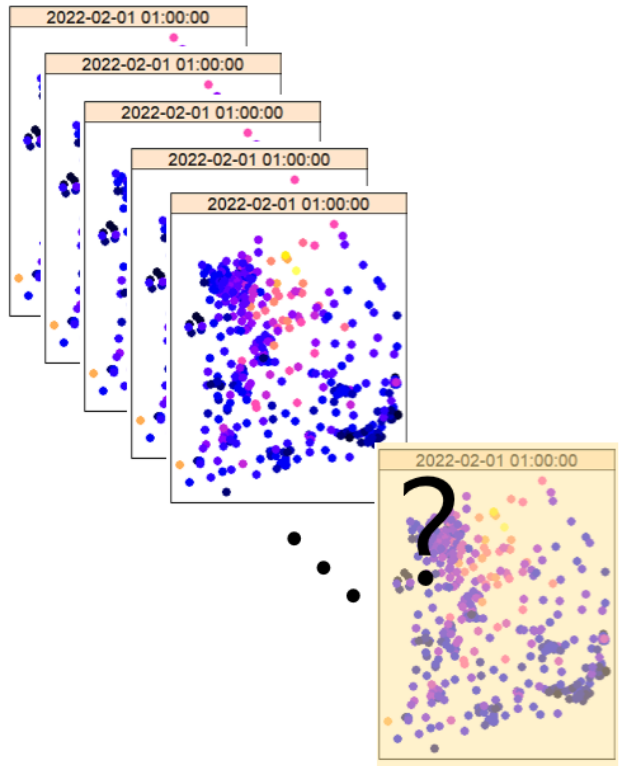
1. 개요

프로젝트 목표

1.



2.



1. 시공간 데이터 예측에서 좋은 성능을 보인다.
2. 시공간 데이터 미래 예측에서 좋은 성능을 보인다.

Section 2

시공간 데이터 예측

2. 시공간 데이터 예측

모델 구조

- 기본 MLP 네트워크 구조의 목표는 어떠한 함수 f에 근사하는 것이라는 사실에 기반한다.
- Universal Approximation Theorem라는 이론적 근거를 통해 MLP 네트워크는 어떠한 함수든 근사할 수 있으며 이를 통해 회귀(Regression) 예측이 가능하다.
- 따라서 가우시안 가정과 정상성 가정이 성립하지 않는 복잡한 데이터에 대해서도 신경망은 이를 학습하고 좋은 예측 성능을 보일 것이다.

2. 시공간 데이터 예측

모델 구조

- 데이터의 분포인 시공간 확률과정(Spatio-temporal random process)

$$Y(\mathbf{s};t) = \mathbf{x}(\mathbf{s};t)^\top \boldsymbol{\beta} + \eta(\mathbf{s};t)$$

에 대해서 시공간 정보의 차원을 늘리기 위해 Cressie (2019)의 사실을 이용한다.

- 충분히 큰 $K \in \mathbb{N}$에 대해

$$\begin{aligned}\hat{Y}(\mathbf{s};t) &= \mathbf{x}(\mathbf{s};t)^\top \hat{\boldsymbol{\beta}} + \hat{\eta}(\mathbf{s};t) \\ &= \mathbf{x}(\mathbf{s};t)^\top \hat{\boldsymbol{\beta}} + \sum_{k=1}^{K} w_k \phi_k(\mathbf{s};t).\end{aligned}$$

으로 다시 쓸 수 있다.

2. 시공간 데이터 예측

모델 구조

$$\left(\mathbf{x}(\mathbf{s};t)^\top, \mathbf{s}, t\right)^\top \in \mathbb{R}^{p+d+1}$$

$$\Downarrow$$

$$\left(\mathbf{x}(\mathbf{s};t)^\top, \phi_1(\mathbf{s};t), \cdots, \phi_K(\mathbf{s};t)\right)^\top \in \mathbb{R}^{p+K}$$

2. 시공간 데이터 예측

모델 구조

- 확장된 입력:

$$\left(\mathbf{x}(\mathbf{s};t)^\top, \phi_1(\mathbf{s};t), \cdots, \phi_1(\mathbf{s};t)\right)^\top,$$

$$\Downarrow$$

  - **입력1**: $\mathbf{x}_1(\mathbf{s};t) = \mathbf{x}(\mathbf{s};t) \in \mathbb{R}^p$
  - **입력2**: $\mathbf{x}_2(\mathbf{s};t) = (\phi_1(\mathbf{s};t), \cdots, \phi_K(\mathbf{s};t))^\top \in \mathbb{R}^K$

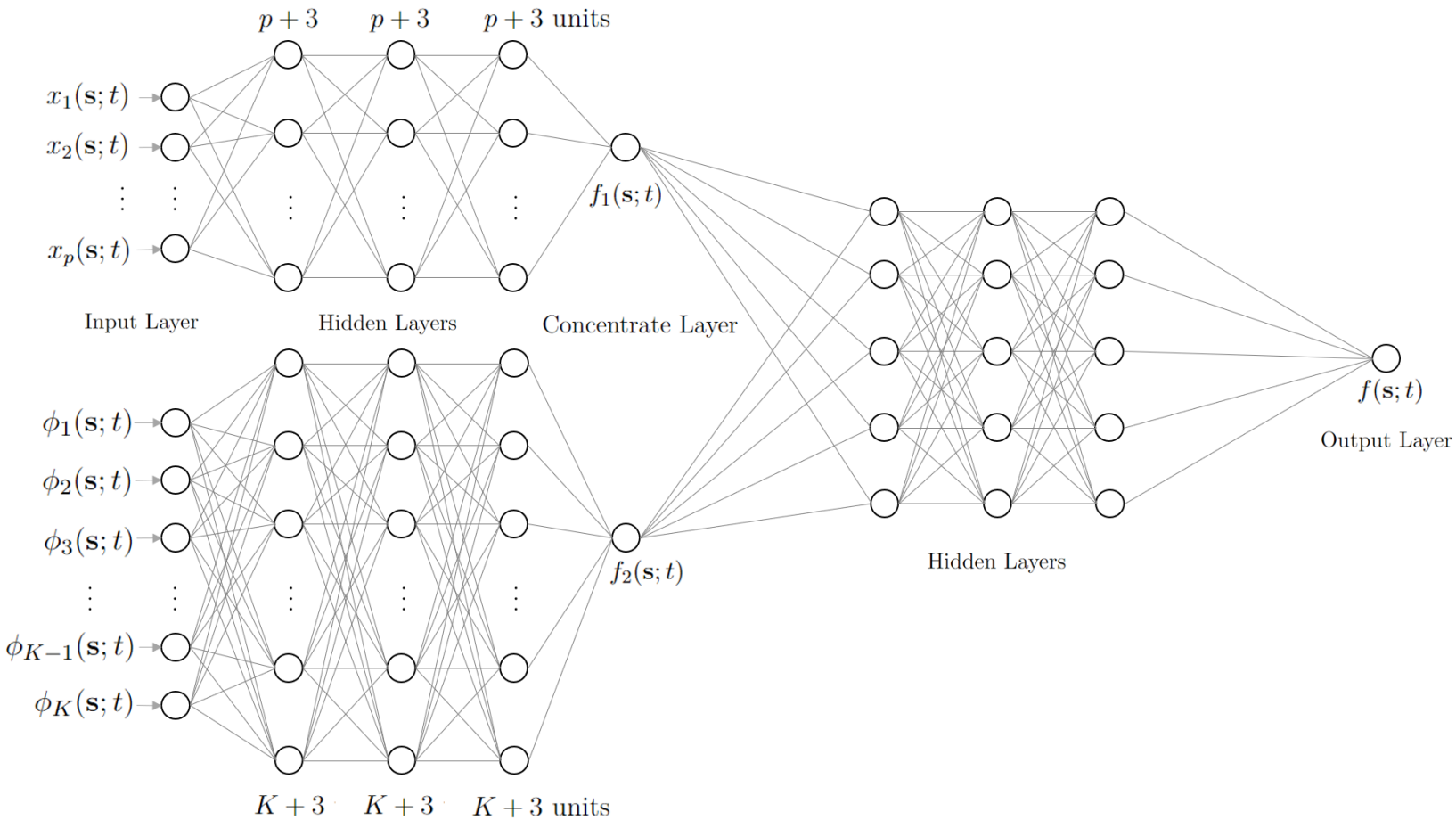
- 각각의 입력에 대해 따로 학습한 출력값을 얻은 후에 앙상블 형태로 최종 출력을 얻는다.

$$\mathbf{z} = \mathbf{x}^\top \boldsymbol{\beta} + \varepsilon_1$$

$$\mathbf{z} = \boldsymbol{\eta} + \varepsilon_2$$

2. 시공간 데이터 예측

모델 구조



2. 시공간 데이터 예측

모델 구조

Definition (Excess Risk)

The **excess risk**, $\mathcal{E}(f)$, compares the risk of f to the $\hat{f}_{Bayes}$:

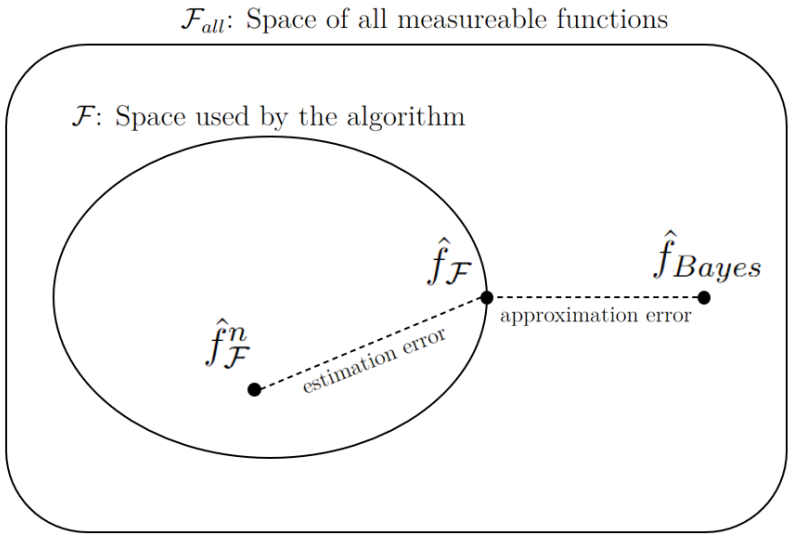
$$\mathcal{E}(f) := R(f) - R(\hat{f}_{Bayes})$$

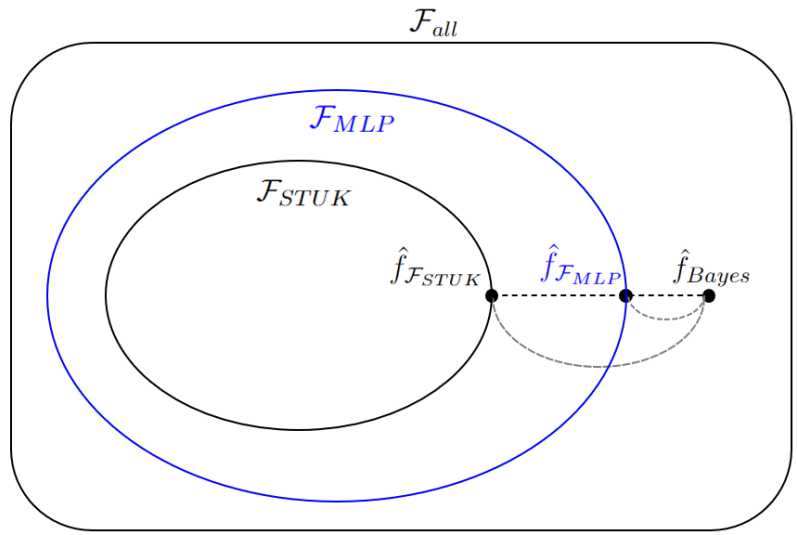
- The excess risk of the $\hat{f}_{\mathcal{F}}^n$ can be decomposed:

$$\begin{aligned}\mathcal{E}(\hat{f}_{\mathcal{F}}^n) &= R(\hat{f}_{\mathcal{F}}^n) - R(\hat{f}_{Bayes}) \\ &= \underbrace{R(\hat{f}_{\mathcal{F}}^n) - R(\hat{f}_{\mathcal{F}})}_{\text{estimation error}} + \underbrace{R(\hat{f}_{\mathcal{F}}) - R(\hat{f}_{Bayes})}_{\text{approximation error}}\end{aligned}$$

2. 시공간 데이터 예측

모델 구조





$$R(\hat{f}_{\mathcal{F}_{MLP}}) < R(\hat{f}_{\mathcal{F}_{STUK}}) \Leftrightarrow \inf_{f \in \mathcal{F}_{MLP}} R(f) < \inf_{f \in \mathcal{F}_{STUK}} R(f)$$

- Universal Approximation Theorem, Kidger et al. (July 2020)를 이용.

2. 시공간 데이터 예측

실제 데이터 실험 결과

- 예측 오차를 알아보기 위해 실험을 한 데이터는 우리나라의 미세먼지($PM_{2.5}$) 데이터이다.
- 미세먼지의 농도에 따라 총 3일의 데이터셋을 이용해 실험하였다.

1. 2022년 2월 1일
   - 446개의 관측소, 24시간, 총 10704개의 관측값
   - 미세먼지 평균 농도 $=$ 약 $29.008\mu/m^3$
2. 2022년 1월 9일
   - 481개의 관측소, 24시간, 총 11544개의 관측값
   - 미세먼지 평균 농도 $=$ 약 $67.419\mu/m^3$
3. 2022년 6월 29일
   - 425개의 관측소, 24시간, 총 10200개의 관측값
   - 미세먼지 평균 농도 $=$ 약 $5.153\mu/m^3$

2. 시공간 데이터 예측

실제 데이터 실험 결과

- Covariate: 기온, 바람, 바람의 x축 성분, 바람의 y축 성분, 강수량, 습도 (6개)
- 실제적인 적용 상황에서는 Overfitting을 줄일 수 있는 Dropout기법과 모델을 안정화 할 수 있는 Batch Normalization 기법을 적용.
- 은닉층의 개수는 3으로 설정.
- 활성함수는 ReLU 함수, 출력함수는 Linear 함수를 사용.
- GPU: GCP의 1xNVIDIA-TESLA-V100 GPU

2. 시공간 데이터 예측

실제 데이터 실험 결과

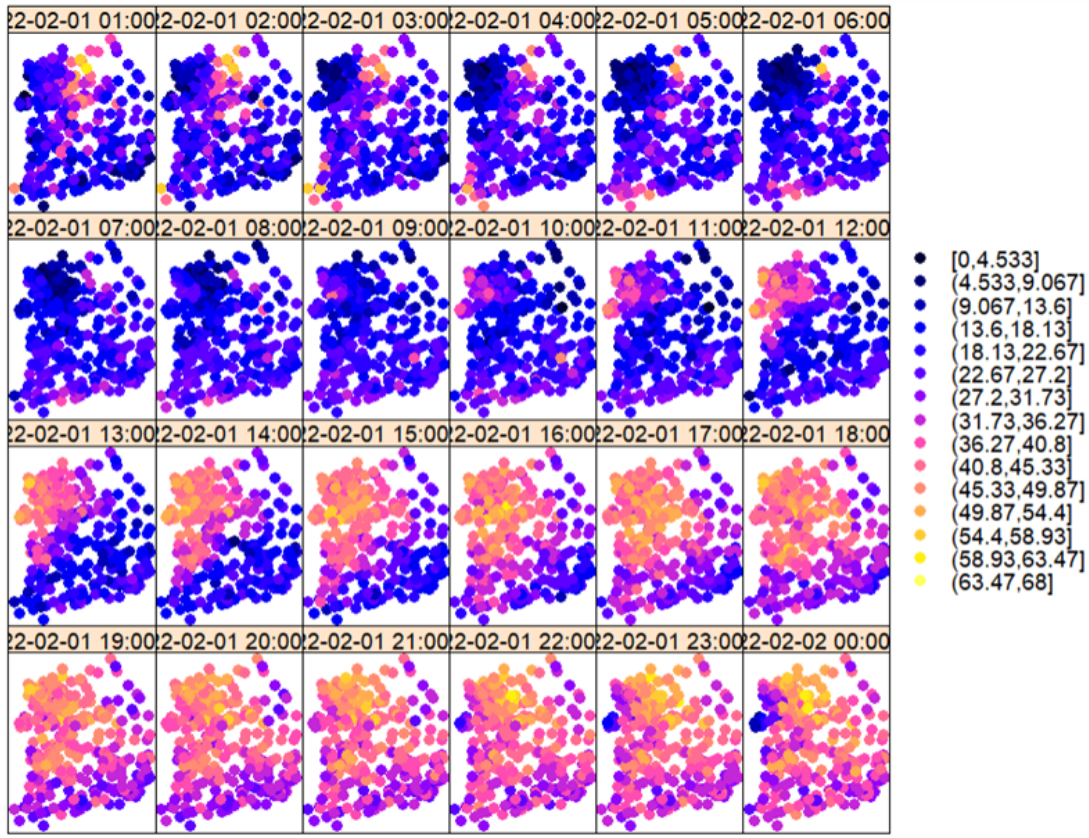
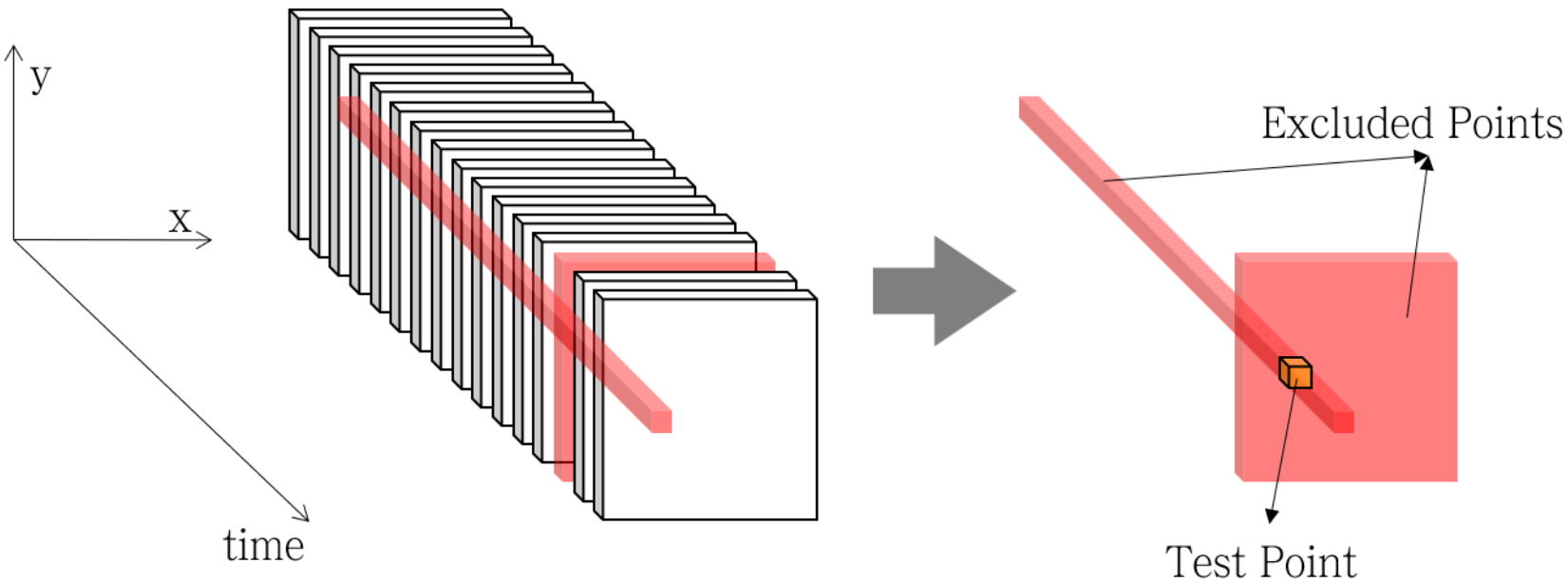


Figure: 2022년 2월 1일의 미세먼지 데이터

2. 시공간 데이터 예측

실제 데이터 실험 결과

- 검정 방법은 다음과 같다.



2. 시공간 데이터 예측

실제 데이터 실험 결과

- 검정 결과는 다음과 같다.

		STUK	Ours	Method1
2022.02.01.	MSE	65.416	**41.488**	51.995
	sd	18.889	14.646	20.575
2022.01.09.	MSE	405.095	**105.832**	124.278
	sd	111.846	36.108	32.412
2022.06.29.	MSE	13.616	**12.402**	13.372
	sd	3.170	4.937	4.552

Section 3

시공간 데이터 미래 예측

3. 시공간 데이터 미래 예측

모델 구조

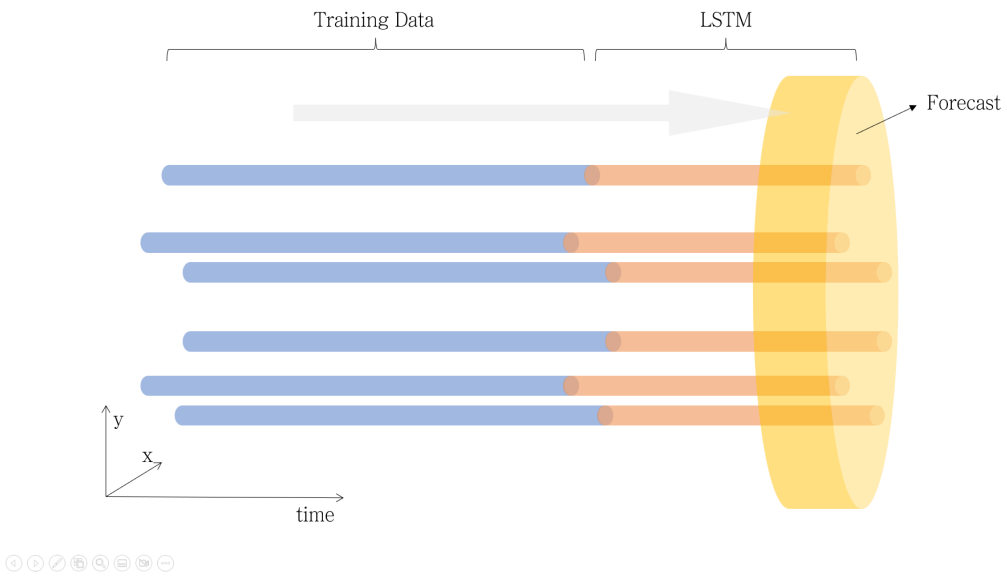
LSTM

- LSTM은 이전 단계의 결과값이 다음 단계의 입력값과 연결된 구조를 가지고 있어 순차 데이터의 분석에 적합하다.
- LSTM은 RNN의 장기 의존성 문제를 보완한 기법으로, 길이가 긴 순차 데이터도 잘 처리할 수 있다.
- 따라서 LSTM을 사용하면 장기간에 걸쳐서 관측된 시공간 데이터에 대해 좋은 예측 성능을 보일 것이다.

3. 시공간 데이터 미래 예측

모델 구조

LSTM



- 각 공간별로 LSTM을 사용해 시간에 따른 패턴을 파악하여 미래의 값을 예측한다.
- 예측된 값들의 공간적 의존성에 기반해서 관측치가 없는 위치의 미래의 값을 예측하려 한다.

3. 시공간 데이터 미래 예측

모델 구조

Spatial Kriging

- Spatial Kriging은 관측되지 않은 위치의 값을 관측된 데이터의 가중 평균으로 예측하는 기법으로, 이 때 가중치는 공간적 의존성을 반영한다.

- 따라서 시간이 고정된 시공간 데이터에 대해 Spatial Kriging으로 예측을 수행할 수 있다.
    - Ordinary Kriging 기법 사용.

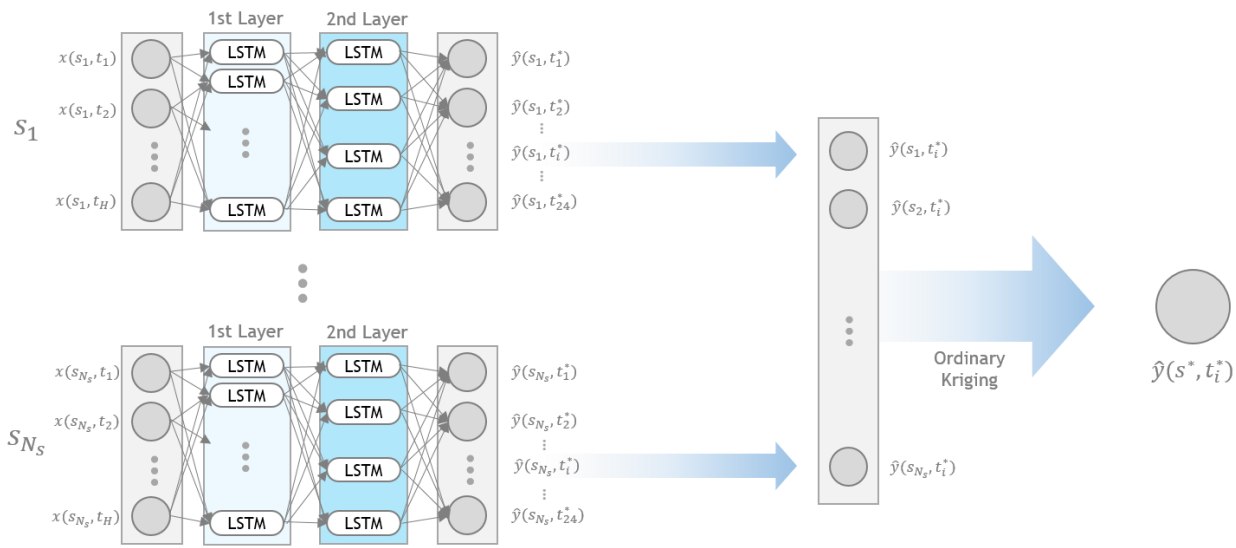
3. 시공간 데이터 미래 예측

모델 구조

LSTM + OK Architecture

- LSTM Input: $X(s, \mathbf{t}) \in \mathbb{R}^{\text{History size} \times (p+1)}$
- LSTM Output: $\hat{Y}(s, \mathbf{t}^*) \in \mathbb{R}^{\text{Target size} \times 1}$
- Ordinary Kriging Input: $(\hat{y}(\mathbf{s}, t^*), lat_{\mathbf{s}}, long_{\mathbf{s}}) \in \mathbb{R}^{N_s \times 3}$
- Ordinary Kriging Output: $\hat{y}(s^*, t^*) \in \mathbb{R}$

  where N_s: # of locations used for ordinary kriging



3. 시공간 데이터 미래 예측

실제 데이터 실험 결과

Dataset

- 실험에 사용된 데이터는 우리나라의 종관기상관측(ASOS) 데이터이다.
- Time: 2021-01-01 00:00 $\sim$ 2022-11-31 23:00 (hourly data)
- Location: 80개의 관측소
- Target: 기온
- Covariate: 강수량, 풍속, 풍향, 습도, 증기압

3. 시공간 데이터 미래 예측

실제 데이터 실험 결과

Model Description: LSTM + OK

LSTM

- History size = 120, Target size = 24 로 지정.
- LSTM(8), LSTM(4)의 두 층으로 구성.
- ReLU 활성 함수, Adam 옵티마이저 사용.

Ordinary Kriging

- Matern 배리오그램 모형 사용.
- 모형의 모수는 AWLS(Approximated Weighted Least Squares)로 추정.

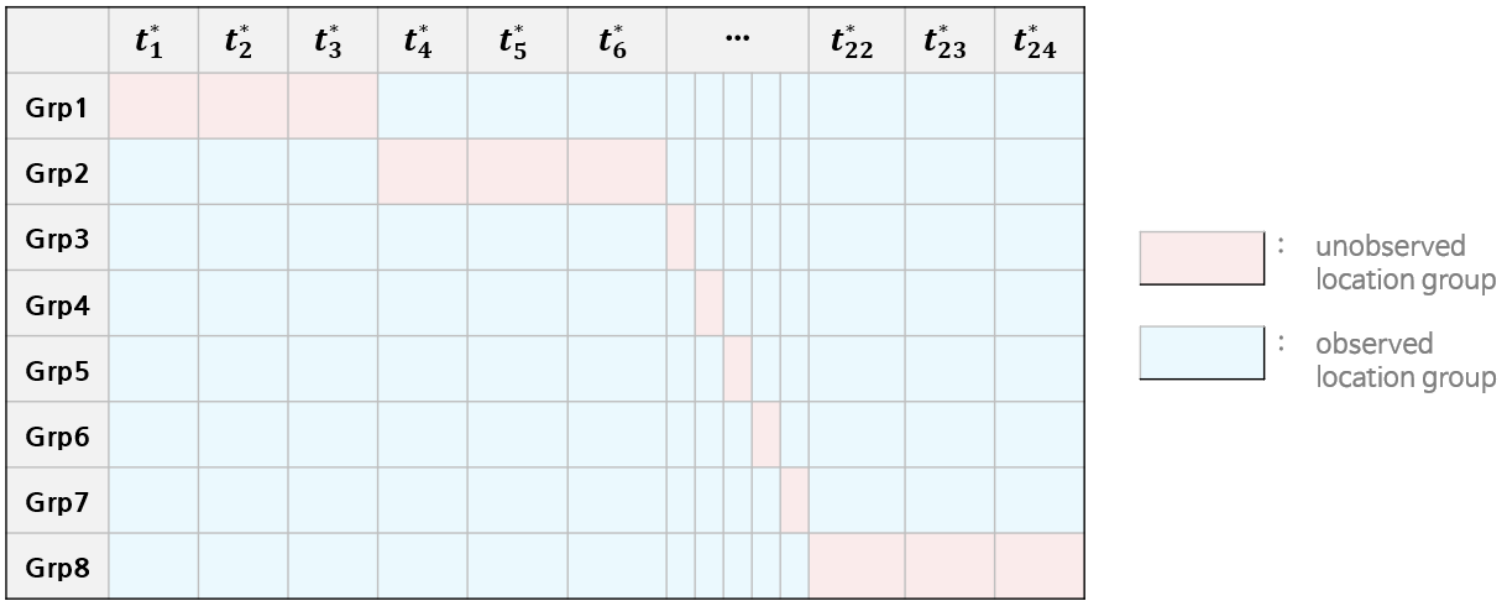
Model Description: Baseline

- Ordinary Spatio-Temporal Kriging 모델 사용.

3. 시공간 데이터 미래 예측

실제 데이터 실험 결과

Evaluation Procedure



- 80개의 관측소를 임의로 8개의 그룹으로 나누고, 관측치가 없는 그룹 24개 각각에 대해 MSPE 계산한다.

3. 시공간 데이터 미래 예측

실제 데이터 실험 결과

Results

	Baseline	LSTM + OK
mean(MSPE)	83.38433	61.85546
sd(MSPE)	25.51655	25.90973

- LSTM + OK 모델의 평균적인 MSPE가 Baseline 모델보다 작게 나오는 것을 확인할 수 있다.

Section 4

결론

4. 결론

- DNN 구조를 활용하여 복잡한 시공간 데이터에서도 좋은 예측 성능을 보이는 모델 제시.
- 실험 결과, 제안한 모델이 시공간 데이터 예측과 시공간 데이터 미래 예측 각각에서 기존 모델보다 좋은 성능을 보임.

References

- 프로젝트 보고서에 수록

감사합니다.