Modeling Central Valley Land Subsidence with Neural Networks

Jay Upadhyay JUPAD002@UCR.EDU
University of California, Riverside

Abstract

The rate at which land is subsiding poses many threats to the environment and infrastructure. Several methods have been trying to capture subsidence accurately. We explore several models including, LSTM, CNN-LSTM, BiDirectional LSTM, and Random Forest using multiple variables such as subsidence, precipitation,and groundwater depth.
Keywords: Multivariate Time Series, Neural Network, Land Subsidence

1. Introduction

Land Subsidence threatens land and infrastructure deformation, flood protection, agricultural production. The wide impacts means proper countermeasures are needed to address the issue. Predict, Prevent, and Protect are key stages in mitigating the disastrous affect to our communities and environment. Prediction is required to identify key areas of heavy subsidence.

From predicting high risk areas, key actions may be taken in prevention of groundwater pumping, mining, and other human activities.

Protection means active intervention to protect these lands from subsidence by means of refilling groundwater wells, and other subsidence reversal efforts.

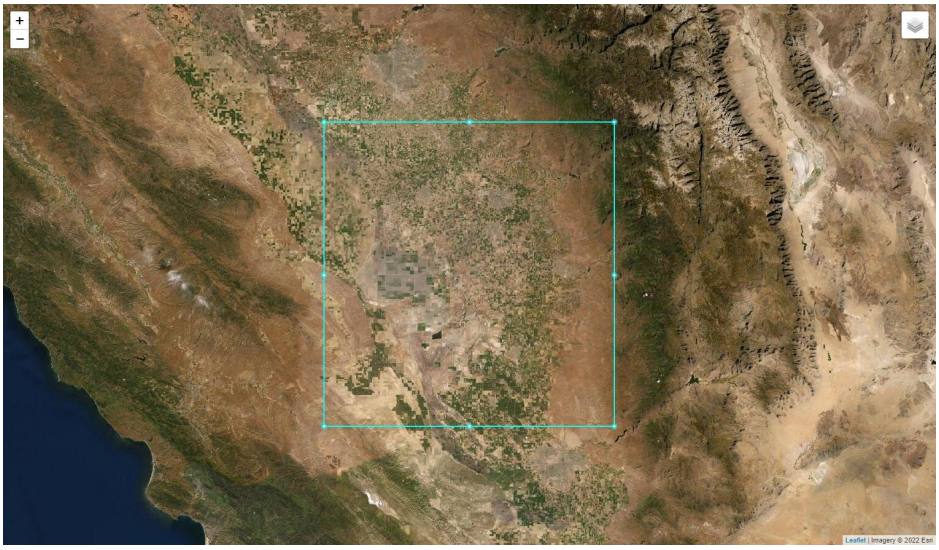
To deal with water demand we are mining groundwater that cannot be readily replaced. Can relate this back to land subsidence.

There is complexity in the underlying mechanisms and effects of subsidence. However, depth to ground water and precipitation are two factors that directly correlate to soil composition and therefore subsidence. Other factors which may be later considered are temperature, elevation.

2. Materials and Methods

2.1. Study Area

The area was California's Central Valley. The area data set spanned from roughly 34 to 37 latitude and -121 to -117 longitude. The image below highlighted the the key area we inspected.



2.2. Methodology

Three main steps were used to accurately predict.

1. Data Handling and Preparation
2. Model Prediction
3. Testing

2.3. Data Handling and Preparation

The displacement data set represented displacement over time. Roughly every two weeks from 2014-2019, we could monitor the subsidence changes through these time series. Given X time series representative over the X area. The data had large amount NaNs and largely consisted of oscillatory patterns indicative of non-subsidence.

The two other features, groundwater data and precipitation followed the same pattern in presentation however consisted of many repeated time series. Another slight issue is they are sampled from a smaller period of 2016-2019. These data sets needed to be slightly cleaned given large amounts of repeated data. Furthermore the longitudes and latitudes did not perfectly align with the displacement data set. We could either interpolate or just focus on the time series that share longitude and latitude between all features. Given time restraints, the decision was made to go with the latter.

For the missing data from 2014, we interpolated groundwater data by including simply filling the missing values with previous or future groundwater. This works exceptionally considering groundwater depth does not radically change often.

For missing precipitation data, we interpolate by gathering yearly data.(Use future January values as missing January value).

We created two data sets to test our models on the monitor performance.

First Dataset The first data set was created by splitting the area into two separate areas of high subsidence into two areas using one for test and the other for training. It was split 60% training,10% validation, 30% testing. Training Set

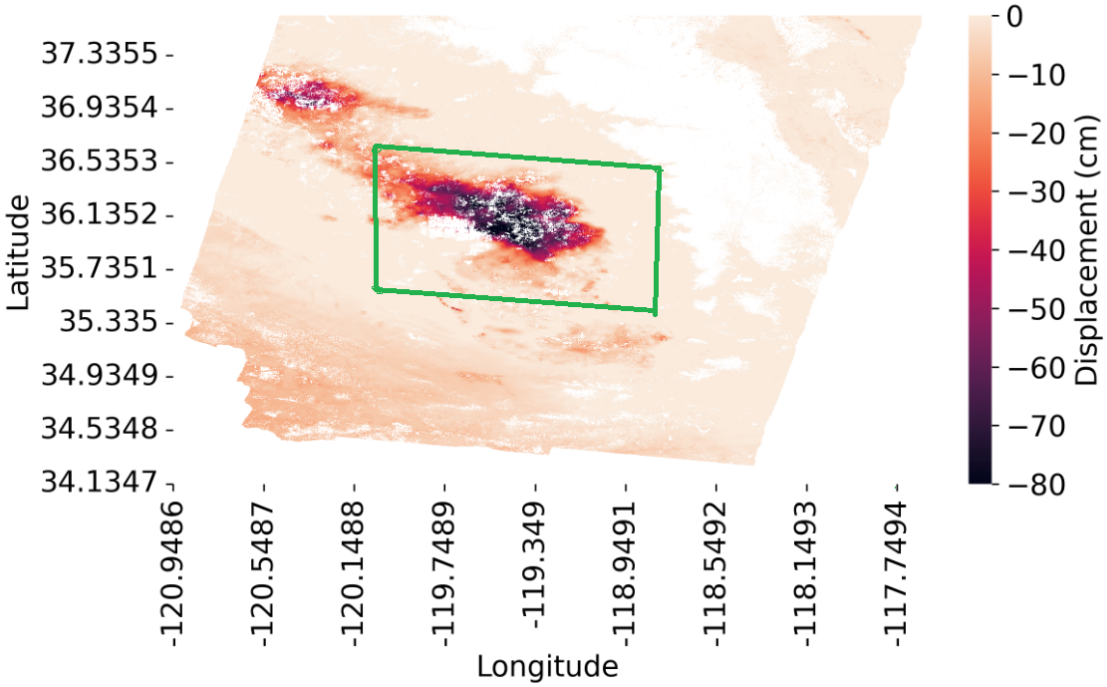
- -118.9000000, 35.5000000
- -118.9000000, 36.5000000
- -119.6000000, 36.5000000
- -119.6000000, 35.5000000

Testing Set

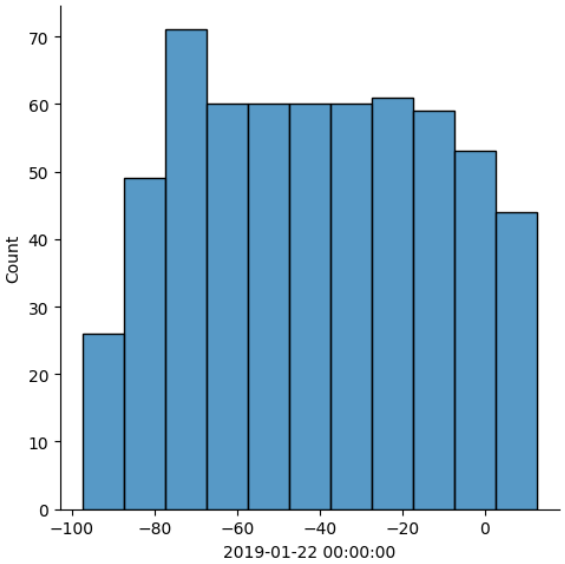
- -118.9000000, 35.5000000
- -118.9000000, 36.5000000
- -120.1000000, 36.5000000
- -120.1000000, 35.5000000

Second data set Similar to the first data set, we took a polygon using the following points.

- -118.8000000, 35.5000000
- -118.8000000, 36.6000000
- -120.1000000, 36.6000000
- -120.1000000, 35.5000000



We then reduced the data by keeping an even distribution across displacements.

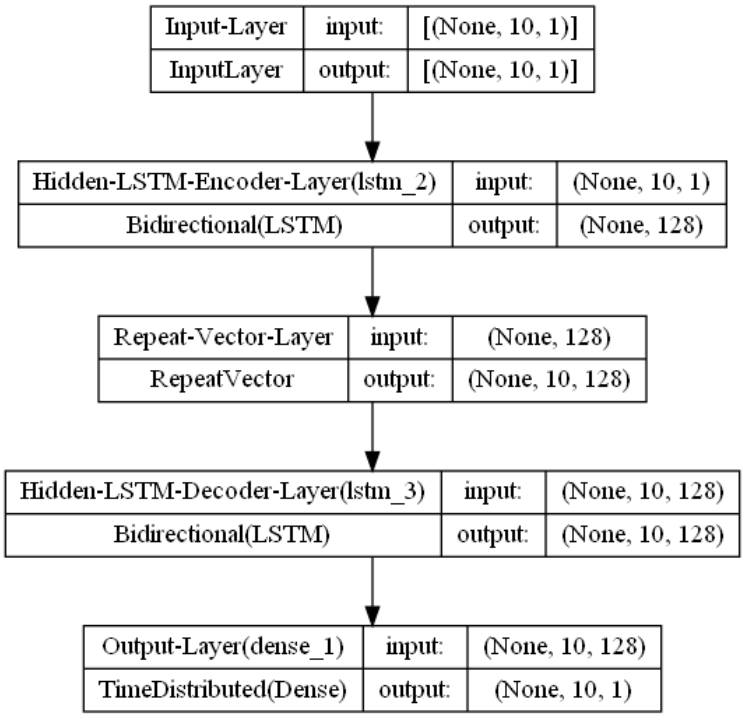


Finally, we shuffled and split 60% of data for training, 20% for validation, and 20% for testing.

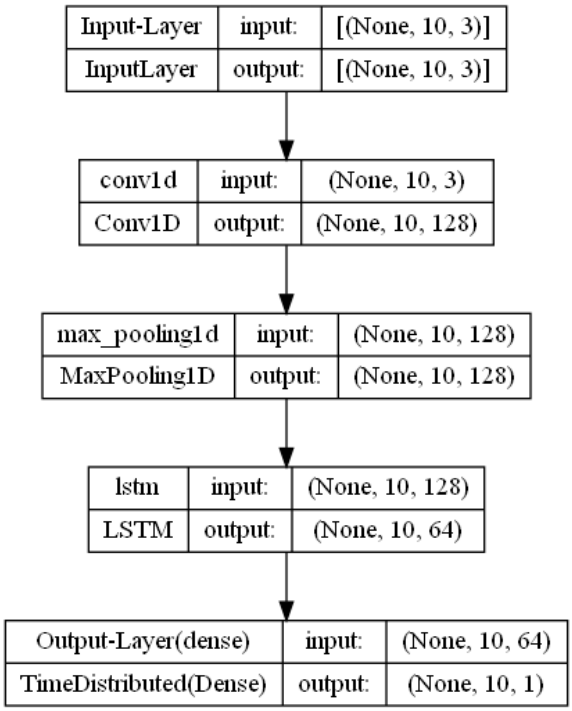
2.4. Model Prediction

We explored and tested different forecasting models with a large interest in LSTMs and Random Forest. LSTM, CNN-LSTM, Bidirectional LSTM all ran with adam optimizer, 100 epochs, .001 lr, batchsize=32. The random forest ran with 100 estimators.

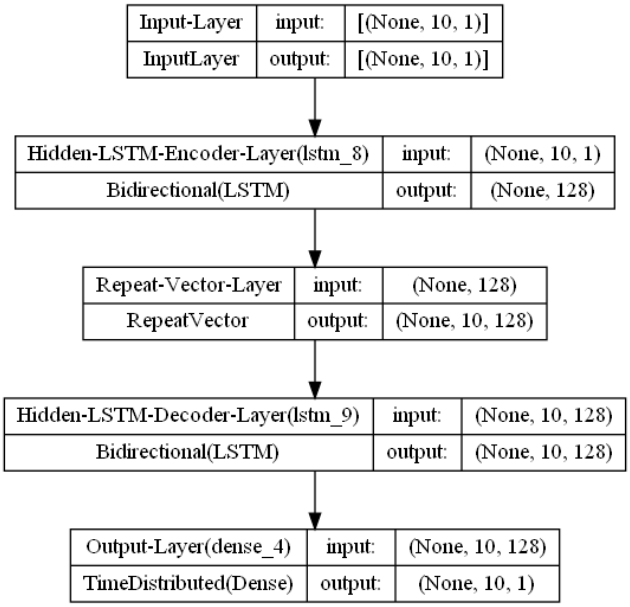
LSTM The LSTM follows the following model diagram. It is relatively simplistic and offered a good beginning benchmark to consider other solutions by.



CNN-LSTM The idea gathered from X paper was to use the CNN to extract the features of the data and use the LSTM portion to predict. There has been shown efficacy in field of stock prices however it did not seem to work for our dataset. Although it has shown to perform worse for not stationary series, there is still room for optimization for this model.



BiDirectional-LSTM The Bidirectional LSTM has become one of our best performing models. It excels at sequence to sequence prediction which is why we believe it performed exceptionally.



Random Forest There have been previous proof of efficacy of random forest and other tree-based methods(Include references). Random forest being one of the best performing models in research left us with extremely high hopes. Our implementation was running a random forest regressor with 100 estimators.

2.5. Testing

The methodology towards testing was to pick certain areas representative of every type of subsidence. By testing the model inclusive of the range of expected subsidence, the results are indicative of performing in both subsiding and non-subsiding areas.

3. Results

Results				
	Dataset 1		Dataset 2	
Metrics / Methods	MAE	MSE	MAE	MSE
LSTM(Univariate Unscaled)	3.06	16.44	2.29	11.85
LSTM(Multivariate Unscaled)	2.10	19.16	1.49	4.68
CNN-LSTM(Multivariate Unscaled)	2.15	8.29	1.44	4.27
BIDIRECTIONAL LSTM(Univariate Unscaled)	1.78	5.98	0.81	1.34
BIDIRECTIONAL LSTM(Univariate Scaled)	1.78	6.04	1.57	4.16
BIDIRECTIONAL LSTM(Multivariate Unscaled)	1.60	5.34	0.82	1.16
BIDIRECTIONAL LSTM(Multivariate Scaled)	1.94	6.30	1.23	2.65
Random Forest(Univariate)	1.73	5.60	1.25	3.40
Random Forest(Multivariate)	1.55	4.26	0.82	1.43

4. Conclusion

In this paper, we compared different types of univariate and multivariate models to determine best forecasting methods for subsidence. I excluded the scaled models from these final rankings because they always performed worse than the unscaled versions. The result for both models followed this order best to worse:

First dataset:

1. Random Forest Multivariate
2. BiLSTM Multivariate
3. Random Forest Univariate
4. BiLSTM Univariate
5. CNN-LSTM
6. LSTM(Univariate)
7. LSTM(Multivariate)

Second dataset:

1. BiLSTM(Multivariate)
2. BiLSTM(Univariate)
3. Random Forest(Multivariate)
4. Random Forest(Univariate)
5. CNN-LSTM
6. LSTM(Multivariate)
7. LSTM(Univariate)

The models performed as expected with the exception of Univariate LSTM outperforming Multivariate LSTM in first dataset. In the second dataset, this anomaly does not appear indicating it may be fault of testing methods.The second exception was Bi-LSTM outperforming Random Forest models entirely on the second dataset.

5. Discussion

The majority of loss seemed to come from volatile non-subsiding series. Certain time series vary heavily which makes it extremely unpredictable. The best of the models succeeded in predicting subsiding series which was the main goal.

The first data set may be faulty to reasonably extract conclusions from. It is anomalous with the univariate LSTM outperforming the multivariate LSTM.

There is an indication that a BiDirectional LSTM may be able to edge out Random Forest in performance however requires more rigorous testing. Furthermore, given the performance of the BiDirectional LSTM, a CNN-Bidirectional LSTM may have some merit. Considerations towards trying to train and predict models on different areas of the world may be suitable given availability of data although there will likely be a performance loss.

The models performance heavily relies on proper training and testing therefore these methods should scrutinized and evolved to better benchmark these models. Upon creating more protection for unreplenishable groundwater wells, it should be considered how it will affect California's water supply and how we can mitigate the ongoing drought and diminishing water supplies.

6. Next Steps

- Fix Precipitation interpolations
- Interpolate groundwater and precipitation to match subsidence areas
- Explore batch normalization vs regular normalization
- Consider possibility of dropout layer(not likely)
- Consider effects of differencing time series data.
- CNN-BiDirectional LSTM

7. Citations and Bibliography

Acknowledgments

We thank Mariam Salloum, Kyongsik Yun, Analisa Flores.

References

[1] *Wenjie Lu, Jiazheng Li, Yifan Li, Aijun Sun, Jingyang Wang, "A CNN-LSTM-Based Model to Forecast Stock Prices", Complexity, vol. 2020, Article ID 6622927, 10 pages, 2020. https://doi.org/10.1155/2020/6622927*, https://www.hindawi.com/journals/complexity/2020/6622927/

[2] *Omid Rahmati, Fatemeh Falah, Seyed Amir Naghibi, Trent Biggs, Milad Soltani, Ravinesh C. Deo, Artemi Cerdà, Farnoush Mohammadi, Dieu Tien Bui, Land subsidence modelling using tree-based machine learning algorithms, Science of The Total Environment,Volume 672,2019,Pages 239-252,ISSN 0048-9697, https://doi.org/10.1016/j.scitotenv.2019.03.496.*, https://www.sciencedirect.com/science/article/pii/S0048969719315128