

DOCTORAL THESIS

Learning Complex Spatio-Temporal Dependency: Model Design, Information-Theoretic Analysis, and Systematic Validation

TAN, Qi

Date of Award:
2021

[Link to publication](#)

General rights

Copyright and intellectual property rights for the publications made accessible in HKBU Scholars are retained by the authors and/or other copyright owners. In addition to the restrictions prescribed by the Copyright Ordinance of Hong Kong, all users and readers must also observe the following terms of use:

- Users may download and print one copy of any publication from HKBU Scholars for the purpose of private study or research
- Users cannot further distribute the material or use it for any profit-making activity or commercial gain
- To share publications in HKBU Scholars with others, users are welcome to freely distribute the permanent URL assigned to the publication

Abstract

Learning the intrinsic dependency among heterogeneous data is one of the most challenging problems in predictive spatio-temporal analytics (PSTA) tasks because the dependency to be learned generally manifests, interacts, and integrates in a complex way at varying scales within and/or across diverse data sources, which renders characterization of such underlying dependency extremely difficult. To address this challenge, a key question must be answered: Given a specific learning task and the corresponding dataset, how should one design an appropriate learning model and theoretically determine its desired configuration by analyzing its learning behavior and quantifying its learning capacity so that the useful information contained in the data can be effectively extracted to yield outstanding predictive performance?

Existing methods, including classical time series models, representative tensor-based learning models, and state-of-the-art deep neural network models, have shown their effectiveness in capturing specific types of dependency, such as short-range temporal dependency or homogeneous/aligned data dependency. However, these methods do not explicitly address the critical issue of multi-scale dependency modeling for heterogeneous data, and their performance thus presents certain limitations. More importantly, even though the existing deep learning models have achieved state-of-the-art performance in certain learning tasks, and earlier studies have examined the behaviors of learning models from the perspective of optimization or representation learning, they are incomplete. Specifically, they failed to provide an in-depth understanding of why a specific deep learning model works well on a given dataset and of the relationship between the learning model's configura-

tion and its corresponding model capacity on the given task and dataset. Without such a clear understanding, it remains a mystery how to determine the desired configurations of a certain deep model with respect to the given learning data sets, thus making it difficult to sustain the success of deep learning.

To address these unsolved yet challenging issues in a theoretically sound and practically feasible way, in this thesis, we aim to systematically answer four research questions:

1. How can we design a learning model to characterize the complex multi-scale dependency of heterogeneous data in PSTA?
2. How can we analyze and quantify the designed model's capacity to capture the multiscale dependency of spatiotemporal data?
3. How can we validate the learning behavior and performance of the designed model in various scenarios of multi-scale spatio-temporal dependency?
4. How can we make the designed model practically applicable?

To answer the first question, we develop and demonstrate a novel interactively and integratively connected deep recurrent neural network (I^2 DRNN) model to capture the multi-scale dependency in heterogeneous spatio-temporal data. The proposed I^2 DRNN comprises three key modules: (i) an Input module that integrates data from heterogeneous sources; (ii) a Hidden module that captures information on various scales while allowing the information to flow between layers in an interactive manner; and (iii) an Output module that models the integrative effects of information from various hidden layers to generate the output predictions. By integrating these modules, the I^2 DRNN can model the integrative effects of various scales of spatio-temporal data within and/or across diverse factors, as observed from heterogeneous sources.

To answer the second question, we propose an information-theoretic framework. This framework enables us to theoretically analyze the learning behavior of I^2 DRNN, quantitatively characterize the information-based model capacity (i-CAP) of each of its compo-

nents in terms of capturing the multi-scale spatiotemporal information, and appropriately determine its necessary and sufficient configurations with respect to a given dataset. With these information-theoretic guarantees, the developed framework serves as rigorous and explainable guidance in the design of a desirable deep architecture for a given task.

To answer the third question, we conduct a series of experiments with both synthetic datasets and real-world PSTA tasks to validate the I^2 DRNN model and to confirm its information-theoretically described behavior. The experimental results show that the I^2 DRNN model outperforms both classical and state-of-the-art models on all datasets and PSTA tasks. Furthermore, as readily validated, the proposed model captures multi-scale spatio-temporal dependency, which is meaningful in the real-world context. More importantly, the model configuration that corresponds to the best performance on a given dataset always falls within the range between the necessary and sufficient configurations, as described by the information-theoretic analysis.

To answer the fourth question, we tackle two practical issues in PSTA: incomplete data imputation before multi-scale dependency learning and hidden interaction mining after multi-scale dependency learning. To infer the missing data for PSTA, we propose a heterogeneous neural metric learning method to restore the data integrity by referring to heterogeneous data sources. We examine the proposed method in a real-world spatio-temporal analytics task, malaria risk mapping, and show that the proposed method produces accurate and reliable risk inference. To characterize the hidden interaction among the spatio-temporal data, we develop two structure-aware methods to uncover the primitive motif prior and the mesoscale connection structure of the underlying interaction network, respectively. Our results with various datasets validate the effectiveness of the proposed methods in capturing the hidden interactions for PSTA.

Keywords: Multi-Scale Spatio-Temporal Dependency Learning, Predictive Spatio-Temporal Analytics (PSTA), Interactively- and Integratively-Connected Deep Recurrent Neural Network (I^2 DRNN), Information-Theoretic Analysis, Incomplete Data Imputation, Hidden Interaction Characterization

Table of Contents

DECLARATION	i
Abstract	ii
Acknowledgements	v
Table of Contents	vii
List of Tables	xi
List of Figures	xiii
Chapter 1 Introduction	1
1.1 Background	1
1.2 Motivations and Objectives	3
1.2.1 Model Design for Multi-Scale Dependency Learning	5
1.2.2 Theoretical Characterization of Learning Capacity	5
1.2.3 Systematic Validation of Learning Behavior and Performance	6
1.2.4 Practical Issues in Model Deployment	6
1.3 Contributions and Significance	7
1.3.1 Multi-Scale Deep RNN Model with Heterogeneous Data	7
1.3.2 Information-Theoretic Framework for Model Analysis	8
1.3.3 Comprehensive Evaluation on PSTA Tasks	8
1.3.4 Tackling Incomplete Data and Hidden Interaction	8
1.4 Structure of the Thesis	9

Chapter 2	Literature Review	11
2.1	Learning Multi-Scale Spatio-Temporal Dependency	11
2.1.1	Predictive Spatio-Temporal Analytics	12
2.1.2	Recurrent Neural Networks for PSTA	13
2.1.3	Theoretical Understanding of Deep Learning	14
2.2	Tackling Practical Issues in PSTA	16
2.2.1	Data Incompleteness	16
2.2.2	Hidden Interaction	18
2.3	Summary	19
I	Multi-Scale Spatio-Temporal Dependency Learning	21
Chapter 3	Interactively- and Integratively-Connected DRNN Model Design	22
3.1	Problem Statement	22
3.2	Preliminaries	23
3.3	I ² DRNN Model	25
3.3.1	Input Module: Integration of Heterogeneous Data Sources	25
3.3.2	Hidden Module: Hierarchical Structure for Multi-Scale Informa- tion Interaction	27
3.3.3	Output Module: Integrative Effects at Multiple Scales	28
3.3.4	The Learning Procedure	30
3.4	Summary	30
Chapter 4	Information-Theoretic Framework for Capacity Characterization	31
4.1	Information-Theoretic Perspective on PSTA	31
4.2	Information-Theoretic Analysis of Capturing Multi-Scale Dependency	32
4.2.1	Learning Capacity of RNN	32
4.2.2	Learning Capacity of I ² DRNN	35
4.2.3	Determination of Model Configuration	38
4.3	Summary	45

Chapter 5	Systematic Validation on PSTA Tasks	46
5.1	Learning Performance Evaluation	46
5.1.1	Synthetic Datasets	46
5.1.2	Real-World PSTA Tasks	51
5.2	Result Analysis and Interpretation	55
5.2.1	Analysis of I ² DRNN's Learning Behavior	55
5.2.2	Interpretation of I ² DRNN's Learning Results	56
5.3	Validation of Necessary and Sufficient Configurations	57
5.3.1	Synthetic Datasets	57
5.3.2	Real-World PSTA Tasks	58
5.4	Summary	58
II	Further Issues in Practical PSTA	62
Chapter 6	Incomplete Data Imputation	63
6.1	Introduction	64
6.2	Problem Statement	66
6.3	Methods	67
6.3.1	Integration of External Factors via HNML	70
6.3.2	Inference on Unobserved Data	73
6.4	Experimental Evaluation	75
6.4.1	Synthetic Datasets	75
6.4.2	Real-World PSTA Task	78
6.5	Summary	90
Chapter 7	Hidden Interaction Mining	92
7.1	Introduction	93
7.2	Problem Statement	95
7.3	Methods	97
7.3.1	Motif-Aware Diffusion Network Inference	97

7.3.2	Mesoscale Anisotropically-Connected Learning	105
7.4	Experimental Evaluation	110
7.4.1	Fine-Scale Network Inference	110
7.4.2	Predictive Analytics with Mesoscale Interaction	120
7.5	Summary	126
Chapter 8	Conclusions and Future work	127
8.1	Conclusions	127
8.2	Future Work	130
	Bibliography	132
	CURRICULUM VITAE	151

List of Tables

3.1	Notations and descriptions.	24
5.1	Overview of the datasets used in three real-world PSTA tasks: I) Disease Prediction, II) Climate Forecast, and III) Traffic Prediction. All tasks include data from heterogeneous sources with different spatial and temporal scales. The target variables are in bold face.	50
5.2	Comparison of the proposed model (I^2DRNN) and six representative models (GP [104], LSTM [48], FS-RNN [90], LRTL [6], ST-ResNet [150], and DA-RNN [101]) on three real-world PSTA tasks: I) Disease Prediction, II) Climate Forecast, and III) Traffic Prediction. The best performance among the seven models on different tasks are highlighted in bold face.	53
5.3	Necessary, sufficient, and the best configurations of I^2DRNN corresponding to different D_s in the synthetic Fractional ARIMA datasets.	59
6.1	Notations and descriptions	68
6.2	Performance evaluation (in terms of MAE) of our method and existing inference methods with different missing patterns and varying missing data rates on the synthetic dataset. The best result for each scenario is underlined and highlighted in bold.	79
6.3	Summary of the underlying risk factors considered in this study, including 2 environmental attributes, 2 geographic attributes, and 22 socioeconomic attributes.	83

6.4	Performance evaluation (in terms of MAE) of our method and existing inference methods with different missing patterns and varying missing data rates on 18 towns in Tengchong with homogeneous features from underlying risk factors. The best result for each scenario is underlined and highlighted in bold.	87
6.5	Performance evaluation (in terms of MAE) of our method and existing inference methods with different missing patterns and varying missing data rates on 62 towns (18 in Tengchong and 44 outside Tengchong) in Yunnan province, with heterogeneous features from underlying risk factors. The best result for each scenario is underlined and highlighted in bold.	88
6.6	Holm’s step-down procedure for analyzing the performance of HNML-UF against other methods.	91
7.1	Notations and descriptions.	98
7.2	Motif counting calculation for seven types of triangle motifs used in motif-aware diffusion network inference framework. $\mathbf{B} = \mathbf{A} \circ \mathbf{A}^T, \mathbf{U} = \mathbf{A} - \mathbf{B}$, in which \circ indicates the Hadamard (entry-wise) product	98
7.4	$F1$ scores of different diffusion network inference methods in the real-world network experiments	116
7.5	$F1$ scores of different diffusion network inference methods in the real-world cascade experiments.	116
7.6	Running time of MADNI in inferring diffusion networks with different sizes on MemeTracker dataset.	116
7.7	Mean absolute error (MAE) of forecasting on the real-world Malaria dataset. . .	119

List of Figures

- 3.1 The proposed Interactively- and Integratively-connected Deep Recurrent Neural Network (I²DRNN) model. (a) I²DRNN is composed of the Input (I) module, the Hidden (H) module, and the Output (O) module. The encoder and decoder structures in I module integrate data from heterogeneous sources. The hierarchical structure in H module is used to capture multiple spatio-temporal effects on target variable caused by covariates from different data sources by allowing interaction of information among various layers. The integrative effects at varying scales are then modeled in O module to generate the output predictions. (b) \mathbf{x}_t is a vector that represents the data from multiple heterogeneous sources in all different locations at time step t and \mathbf{o}_t is an N -dimensional vector representing the predicted values of target variable in N locations at time t . By extracting the information from \mathbf{x}_t using the hidden layers, i.e, \mathbf{h}_t^1 , \mathbf{h}_t^2 , and \mathbf{h}_t^3 , the spatial dependency of various locations can be captured. Specifically, the target variable in one location can be influenced by the effects from individual locations (e.g., node $\mathbf{h}_{t,7}^1$) or the collective effects from different locations at multiple scales (e.g., node $\mathbf{h}_{t,9}^2$ and node $\mathbf{h}_{t,5}^3$). The hierarchical structure can learn such multi-scale spatial dependency. Note that the hierarchical layers are fully connected (i.e., $\mathbf{x}_t \rightarrow \mathbf{h}_t^1$, $\mathbf{h}_t^1 \rightarrow \mathbf{h}_t^2$, $\mathbf{h}_t^2 \rightarrow \mathbf{h}_t^3$, $\mathbf{h}_t^1 \rightarrow \mathbf{o}_t$, $\mathbf{h}_t^2 \rightarrow \mathbf{o}_t$ and $\mathbf{h}_t^3 \rightarrow \mathbf{o}_t$). Some of the connections are highlighted simply for illustration. 26
- 4.1 Information-theoretic perspective on multi-scale spatio-temporal dependency. (a) MI between the target variable $Y_{t,j}$ and input covariates $X_{t-lag,k}$ in a climate dataset, where j represents the location of (286.1E, 40.7N) marked with a pointer

	and lag indicates the time lag of $X_{t-lag,k}$ with respect to $Y_{t,j}$. The colour intensity in each location k visualizes the value of $\sum_{t=lag}^T I(Y_{t,j}; X_{t-lag,k})/T$, with $lag = 0, 4, \text{ and } 8$ weeks in the top, medium, and bottom layers, respectively. (b) Normalized MI between the target variable \mathbf{y}_t and the lagged input covariates $\mathbf{x}_{t-\tau}$, i.e., $I(\mathbf{y}_t; \mathbf{x}_{t-\tau})/I(\mathbf{y}_t; \mathbf{x}_t)$, with varying time lags τ in a traffic dataset.	43
4.2	The advantage of I ² DRNN's structure over a stacked RNN's structure. (a) Compared to stacked RNN (left), the fully connected output module in I ² DRNN (right) provides shortcuts from $\mathbf{h}_t^1, \mathbf{h}_t^2, \mathbf{h}_t^3$ to \mathbf{y}_t to preserve short-term and long-term memories in different layers. (b) Stacked RNN (left) uses only the information in \mathbf{h}_{t-1}^1 in the previous time step; I ² DRNN (right) uses all available information in $\mathbf{h}_{t-1}^1, \mathbf{h}_{t-1}^2, \mathbf{h}_{t-1}^3$ in the previous time step to control the input to prevent redundant information filling in the memory bank.	44
4.3	Necessary configuration (I_n) and sufficient configuration (I_s) of the designed model for a given dataset. (a) The $I(Z; Y)$ curve. (b) The first-order derivative of $I(Z; Y)$. (c) The second-order derivative of $I(Z; Y)$	44
5.1	Performance of I ² DRNN on synthetic datasets. (a) Illustration of settings of the multi-scale copy task. (b) Performance of I ² DRNN and stacked RNN with fixed parameter setting on the two-scale copy task. (c)(d)(e) Performance of I ² DRNN and stacked RNN with varying $N, S1$ and T_s on the two-scale copy task. (f) Performance of I ² DRNN with different configurations on the two-scale copy task. (g) Performance of I ² DRNN with different configurations on the three-scale copy task.	47
5.2	MI between hidden layers and the input layer in various time lags, $I(h_t^l; X_{t-lag})$, on (a) disease prediction task, (b) climate forecast task, and (c) traffic prediction task. The lower layers tend to capture the shorter dependency, while the upper layers tend to capture longer and coarser dependency.	59
5.3	Multi-scale dependency among various regions on three real-world tasks. (a) Correlations at various scales are calculated using the output weights of our model in different layers. (b) Correlation matrices at various scales on the dis-	

	ease prediction task. (c) Correlation matrices at various scales on the climate forecast task. (d) Correlation matrices at various scales on traffic prediction task (left); Top 5 correlated regions to Xujiahui district at various scales (middle); Spatial distributions of dominant attributes to shape spatial correlations at various scales (right). Ten locations for the most dominant attribute with the largest POI counts in each scale are marked as red points. The attributes that make the greatest contribution at the first, second, and third scales are <i>tourist attraction</i> , <i>real estate</i> , and <i>shopping place</i> , respectively.	60
5.4	Necessary, sufficient, and the best configurations of I ² DRNN on (a) disease prediction task, (b) climate forecast task, and (c) traffic prediction task. <i>Top row</i> : Curves of model configuration vs. learning capacity. The range between the necessary configuration and the sufficient configuration is shaded in green, and the best configuration with respect to the test performance is highlighted as a vertical red line. <i>Middle row</i> : First-order derivative of the configuration-capacity curve. <i>Bottom row</i> : Second-order derivative of the configuration-capacity curve.	61
6.1	Illustration of Heterogeneous Neural Metric Learning (HNML) for inferring the current number of infection cases in unobserved locations from incomplete historical data and heterogeneous data sources. (a) Given the incomplete historical data, HNML integrates different disease-related risk factors from heterogeneous data sources, such as environmental, geographical, and socioeconomic factors, to infer unobserved infection cases in different locations. (b) The detailed structure of the HNML. HNML utilizes disease-related risk factors to learn an embedding for each location; the learned embeddings are used to characterize the correlations between different locations. As an example for illustration here, HNML learns two mappings from two types of feature space, i.e., source 1 + 2 and source 2 + 3, to the common space. HNML then integrates the incomplete historical data and the correlations between different locations to restore the integrity of case reporting data and make inferences for unobserved data via regression. .	69
6.2	Illustration of Heterogeneous Neural Metric Learning with Unknown Factors	

	(HNML-UF) for considering unknown factors in spatial correlation modeling. To further consider the unknown factors that shaping the spatial correlations, an extra representation vector <i>emb</i> , as indicated as the red rectangle, is to be estimated for each location.	74
6.3	Mapping of 18 counties in Yunnan province on the China-Myanmar border. (a) and (b) highlight the area in Yunnan province on the China-Myanmar border, covering 18 counties. (c) enlarges a specific county (Tengchong County, including 18 towns) from (b).	80
6.4	Schema of Integrating Heterogeneous Data Source	83
6.5	Illustration of examples of three missing patterns, i.e., Spatial Missing, Temporal Missing and Spatio-Temporal Missing, in the real-world malaria surveillance at Tengchong, a malaria endemic county in the Yunnan province of China, which borders Myanmar [120]. The data marked in red are observed while those marked in gray are missing. (a) Spatial Missing (S-M), (b) Temporal Missing (T-M), and (c) Spatio-Temporal Missing (ST-M).	87
6.6	The effect of integrating varying underlying risk factors in missing value imputation. If the unknown factor module in our method is turned off, incorporating only the environmental and geographical factors in the model has a comparable performance to incorporating environmental, geographical, and socioeconomic factors; if the unknown factor module is turned on, incorporating the extra socioeconomic factors obviously enhances performance.	88
7.1	Schematic illustration of the proposed motif-aware diffusion network inference (MADNI). (a) An example of the individual level diffusion network. (b) An enlarged part of the diffusion network in Figure 1(a) for algorithm illustration. (c) The observed cascades on the underlying diffusion network shown in Figure 1(b). For each cascade, only the infection time of the influenced nodes (the red nodes) are observed, such as $\mathbf{t}^1 = \{t_1^1, t_2^1, \dots, t_{13}^1\}$. (d) The motif profile is mined from the cascade data by estimating the frequency of various motif patterns in the underlying diffusion network. (e) The underlying diffusion network is in-	

	ferred via regularized learning with motif prior. The mined motif profile and the inferred network are alternately refined until the inferred network converges. (f) The diffusion network will be inferred by the MADNI framework.	99
7.2	The penalty of different regularization functions, used in scale-free network inference, on nodes with different degrees.	106
7.3	Illustration of the proposed MACL. (a) First, the locations are grouped into several clusters based on the spatial adjacency. Then the mesoscale connectivity among the clusters is constructed by linking the clusters with their K closest clusters. Section 7.3.2.1 shows the details of spatial clustering and connecting procedure. (b) During the learning, the features of the data in the same cluster and the state information from the neighborhood clusters are utilized to update the state representation for prediction. The scenario that the state information is forwarded $L = 2$ times is illustrated. Section 7.3.2.2 shows the details of the formulations. (c) In a dynamic environment, when incorporating new covariates in the existing cluster, only prediction modules of the clusters within the distance of 2 need to be updated, i.e., training the modules in the dashed-border box with new covariates via error back-propagation, avoiding the retraining of the entire model. (d) In a dynamic environment, when the new covariates are collected in a new cluster, the mesoscale connections are added and the local update is applied. Section 7.3.2.3 shows the details of incremental learning procedure.	108
7.4	Comparison result. (a) Performance (in terms of Precision and Recall) comparison between MADNI and the baseline NETRATE [105]. MADNI significantly outperforms the baseline method with only one iteration. The performance of MADNI is further improved after each iteration, and the learning procedure quickly converges in only three iterations. (b) $F1$ improvement ratio of MADNI over the baseline method with varying proportion of random edges, ρ . MADNI outperforms the baseline even if the target network is close to a random network ($\rho = 0.9$). When the target network becomes more structured, i.e., ρ becomes smaller, the improvement ratio becomes more significant.	112

7.5	Performance (in terms of $F1$) comparison between MADNI and the baseline NE-TRATE in (a) Exponential and (b) Rayleigh cascades on the synthetic networks with different types of closely connected triangle motifs.	114
7.6	Network statistics. (a) The motif frequency of the real email communication network and the real MemeTracker diffusion network. (b) Community structure of the email network. (c) The degree distribution of the email network.	117
7.7	The dot plot of motif counting matrix for the inferred meta-population malaria diffusion network.	119
7.8	Plot of modular networks and corresponding Q scores of the ground-truth network, the inferred network in first interaction and the inferred network in final iteration. The modularity of inferred network increases after refining by MADNI.	121
7.9	Degree distributions of the ground-truth network, the inferred network with $\log l_1$ norm and the inferred network with fan motif in our framework. The scale-free property of the underlying diffusion network can be mined by MADNI with fan motif. The slope of these three power-law distributions are -1.6572 , -0.9957 and -1.1022 , respectively.	121
7.10	Running time of LSTM in 1,000 back-propagation iterations with varying numbers of hidden units (a) and parameters (b). The number of parameters in 1-layer LSTM is $4h^2 + 4hd$, where h is the number of hidden units and d is the input size.	123
7.11	Comparison results on training efficiency and effectiveness. (a) Comparison of training time between LSTM and the proposed framework in 1,000 back-propagation iterations with varying number of hidden units. (b) Comparison of test performance among the proposed framework, MPNN, and LSTM. In our methods, MACL-L1 exchanges the information among the clusters 1 time and MACL-L2 exchanges the information 2 times.	123
7.12	Experimental result on dynamic environment. (a) The vertical bars indicate the epochs in which one new covariate dataset is inputted. The performance of the sequential update version, after it converges, is comparable to that of the batch update version. (b) New locations are added for a trained model. In the entire	

update version, the entire network is updated after the new location is added; while in the local update version, the partial network is updated. The local update version achieves lower RMSE with shorter training time. 125

8.1 Complete framework for learning complex spatio-temporal dependency developed in this thesis. 130