二国間交流事業共同研究・セミナー
令和 3（2021）年度実施報告書
オンライン開催　2022 年 3 月 2 日（水）～3 月 3 日（木）

空間計量経済学およびイメージ処理の最前線：
欠測データ、因果推論、機械学習
Cutting-Edge Issues
in Spatial Econometrics and Image Processing:
Missing Data, Causal Inference and Machine Learning

目的：

本セミナーは，近年，最先端の情報科学技術分野の成長が著しいインドの先端大学での教育と研究の動きを確認すると共に，その最先端の研究成果を通じて学び合い，協力可能な内容を確認しながら，データサイエンスの発展に資するものとして社会に還元できる項目を明らかにする狙いがある．また，これからの時代が要求する「インフォメーションサイエンス」と「データサイエンス」の融合を新たな目標としている長崎大学情報データ科学部の新たな分野の開拓と社会貢献に大きなインパクトを与える狙いもある．

長崎大学

インドラプラズサ情報工科大学デリー校

インドラプラズサ情報工科大学デリー校（IIIT-Delhi, Indraprastha Institute of Information Technology, Delhi）は，デリー州政府によって 2008 年に新設された IIIT のデリー校である．すでに，Times Higher Education World University Rankings (2022)では 601-800 位にランクインしており，QS World University Ranking (by subject CSE 2021)では 401-450 位にランクインするなど，IT 分野における研究機関として特に注目されている．

【プログラム】

2022 年 3 月 2 日（March 2nd, 2022）

13:00-13:05	Ryuei Nishii (Nagasaki University): Past, Present, and Future of School of Information and Data Sciences at Nagasaki University
13:05-13:10	Gaurav Arora (IIIT-D): Past, Present, and Future of Indraprastha Institute of Information Technology, Delhi
13:10-13:30	Gaurav Arora (IIIT-D): Econometric Issues due to Missingness Mechanisms and Spatial Autocorrelation Patterns in Groundwater Levels of India's Gangetic Plains
13:30-13:50	A V Subramanyam (IIIT-D) : IN-SAT: A novel land cover classification dataset for Indian Subcontinent
13:50-14:10	Masayoshi Takahashi (Nagasaki University): A novel application of multiple imputation to causal inference
14:10-14:30	Ryuei Nishii (Nagasaki University): SOLAR ACTIVITY IS ONE OF TRIGGERS OF EARTHQUAKES WITH MAGNITUDES LESS THAN 6
14:30-14:40	Break
14:40-15:00	Senya Kiyasu (Nagasaki University): Semi-Supervised Land Cover Classification with Spectral and Spatial Consideration
15:00-15:20	Muthu Subash Kavitha (Nagasaki University): Artificial intelligence bridges the gap between human and machine
15:20-15:40	Byungdug Jun (Nagasaki University): Displacement Model of the 2016 Kumamoto Earthquake using InSAR Technology

2022 年 3 月 3 日（March 3rd, 2022）

13:00-13:20	Masao Ueki (Nagasaki University): Spatio-temporal nonlinear random effects modeling for COVID-19 case counts in Japan
13:20-13:40	Yuta Umezu (Nagasaki University): Non-linear variable selection with ultra high-dimensionality
13:40-14:00	Yu Ichifuji (Nagasaki University): Geographic factor analysis of hotel plan prices using web booking data
14:00-14:20	Norio Setozaki (Nagasaki University): Development of Immersive VR Learning Equipment for Peace Education with Tangible User Interface
14:20-14:30	Break
14:30-14:50	Makoto Fujimura (Nagasaki University): Degradation Scheme by Geometrically-Transformed Image against Average Attack
14:50-15:10	Yuichiro Shibata (Nagasaki University): Real-time signal and image processing with FPGA computing
15:10-15:30	Tomoya Sakai (Nagasaki University): "Low-rank and sparse" everywhere: what if plugged in deep learning?
15:30-16:30	Discussion

セミナー開催責任者：
長崎大学情報データ科学部 高橋将宜
Email: m-takahashi@nagasaki-u.ac.jp
【共同オーガナイザー：Gaurav Arora 助教授（IIIT-Delhi）】

本セミナーは「二国間交流事業　共同研究・セミナー　令和 3(2021)年度」（国立大学法人長崎大学）・（JPJSBP220217703）の支援により開催されるものである.

共同セミナーウェブサイト：
https://www.idsci.nagasaki-u.ac.jp/archives/2496/

【講演要旨】

Econometric Issues due to Missingness Mechanisms and Spatial Autocorrelation Patterns in Groundwater Levels of India's Gangetic Plains

Gaurav Arora

Indraprastha Institute of Information Technology, Delhi

Abstract:
The groundwater level frequently appears as a variable in empirical models in applied economics, sustainability sciences, hydrology and related disciplines. Governmental agencies monitor groundwater levels and these data are an important input for academic research and policy. Missing values can occur in published data due to planned monitoring decisions or defective, inaccessible and dry observation wells. We ask whether missingness is ignorable in the context of groundwater levels' mean estimation and causal inference. We argue that missing values occurring due to "dry wells" are non-random and show that the deletion of missing values due to dry wells leads to right-censoring and a sample selection bias in estimates of the population mean, OLS regression coefficients, and the average treatment effect.

Groundwater levels are spatially auto-correlated meaning that the level at any given location is more similar to those at nearby locations than the locations that are further away. This can be a source of inefficiency in OLS estimation but also provides opportunities for "better" spatial interpolation. We present a theoretical background on relevant geo-statistical models and demonstrate characterization and estimation of spatial auto-correlation in groundwater levels and discuss its impact on statistical inference using simulations. We discuss the key underlying assumption of “spatial stationarity” and other methodological concerns when studying spatial auto-correlation in groundwater levels.

This is a joint work with Saif Ali.

IN-SAT: A novel land cover classification dataset for Indian Subcontinent

A V Subramanyam

Indraprastha Institute of Information Technology, Delhi

Abstract:
Remote sensing through satellite imagery is applied widely for environmental control, urban planning and land cover classification. To this end, supervised deep learning models can fully exploit the potential of satellite images. However, such models require large annotated datasets. Therefore, through this study we present two novel labelled satellite image datasets, namely In-Sat, based on Sentinel-2 Satellite Imagery and Google Earth Imagery covering Indian subcontinent regions. We provide benchmarks and detailed analysis for these datasets using state-of-the-art and our improved remote sensing deep learning models which achieve 91% and 81% overall accuracy in region-wise split setting and 98% and 94% overall accuracy in class-wise split setting for Sentinel-2 data and Google Earth data respectively. We also demonstrate the application of our system for change detection.

This is a joint work with Meet Shah Harshadbhai and Gaurav Arora.

A novel application of multiple imputation to causal inference

Masayoshi Takahashi

Nagasaki University

Abstract:
Causal inference is often regarded as a missing data problem in the potential outcomes framework. Also, multiple imputation is known to be the general approach to missing data. However, multiple imputation is rarely used as a method for causal inference. In the causal inference literature, it is known that the regression discontinuity design (RDD) can be used to estimate the local average treatment effect (LATE) at the cutoff point. This presentation introduces a novel approach to estimating the LATE at the cutoff point by multiply-imputing potential outcomes, which is named multiple imputation regression discontinuity design (MIRDD). Based on the results from Monte Carlo simulations under 112 different settings, each repeated 5,000 times, we show that the MIRDD performs well in terms of bias, root mean squared error, coverage, and interval length compared to the standard RDD methods. We propose to use the MIRDD as a graphical diagnostic tool for the RDD. To implement the proposed method, an easy-to-use software program is also provided.

SOLAR ACTIVITY IS ONE OF TRIGGERS OF EARTHQUAKES WITH MAGNITUDES LESS THAN 6

Ryuei Nishii

Nagasaki University

Abstract:
Possible triggers of earthquakes such as tidal stress, rainfall, the building of artificial water reservoirs, mining, and extraction of fossil fuels have been discussed in the literature. Furthermore, solar activity has been also speculated to be a trigger of earthquakes. In this article, we tackle this problem by data analysis based on machine learning, wherein we used physical measurements of solar activity and related variables for predicting frequencies of earthquakes on the entire earth.

From the analysis of actual data, we conclude that solar activity is one of the triggers of earthquakes with magnitudes from 3.0 to 5.9. In addition, we show that the Earth's electric field is the most effective variable related to earthquakes.

This is a joint work with Pan Qin and Ryosuke Kikuyama.

Semi-Supervised Land Cover Classification with Spectral and Spatial Consideration

Senya Kiyasu

Nagasaki University

Abstract:
Land cover identification is one of the most important purpose of remote sensing. One of the prescribed categories is assigned to each of the pixels according to the spectral characteristics of the land cover objects and the land cover map are produced. The supervised classification is the commonly used method of analysis when the objective categories are prescribed and training data for each category are available. However, it is almost difficult to always provide sufficient data for estimating the spectral characteristics in real applications. This situation often causes considerable errors and difficult to be solved using only the provided training data.

We use the semi-supervised methods for solving this problem which train the classifier using both of the provided training data and the unclassified objective image. We applied the method of semi-supervised land cover classification considering spectral and spatial arrangement in the image.

One of our approaches is as follows. We partition all pixels into k clusters considering both of spectral and spatial characteristics. Then we select several clusters out of them which can be assigned appropriate categories using the provided training data. We use the pixels in the selected clusters as additional training data and classify all pixels in the image by using the sequence of supervised classification. However, the results of clustering sometimes provide inappropriate clusters. We tried to improve the results by assigning the category to the predefined regions in the image which are considered to consist of one category of object.

This is a joint work with Aye Moh Htun and Kotaro Sonoda.

Artificial intelligence bridges the gap between human and machine

Muthu Subash Kavitha

Nagasaki University

Abstract:
Humans have really impressive abilities and can act by both thinking and intuition. How do we bring that knowledge into a more structured prediction model and how inspiration from brain mechanisms could give us good priors to design a good neural network architecture. In order to make convolutional neural networks (CNNs) better, we need to know about how the input is being processed by the brain to object category. If we did understand that then we need to know where we use that piece of understanding. Our brain's plasticity allows us to learn and improve our skills. Similarly neural network learns when we feed it with lots of data. During the training process, the weights are tuned accordingly to strengthen or weaken the connection between neurons. This is how neuroscience inspires artificial intelligence (AI) in emulating human intelligence and building neural networks that mimic brain structure. We proposed specific knowledge to enhance the power of the neural networks in identifying objects and explain briefly two AI projects. In the study for the delineation of the vessel regions on Eye Image calculated from topological characteristics can able to detected damaged tiny vessels, even that do not appear in the ground truth but appear in the original fundus images, which is more appealing on the relevant literature of segmentation. Another study developed generative model which used combined values of both normal and error spaces of autoencoder based on orthogonal projection constraints to detect anomaly objects. The proposed AI models performance is compared over the network without our proposed scheme. The network with our proposed function showed much higher performance than the state-of-the-arts. The effectiveness of the components of AI are evaluated using ablation study experiments.

Displacement Model of the 2016 Kumamoto Earthquake using InSAR Technology

Byungdug Jun

Nagasaki University

Abstract:

Meteorological Agency (2016), the magnitude of a foreshock on April 14 was 6.5 and that of the main shock on April 16 was 7.3. The epicenter was in Mashiki-cho located in the east of Kumamoto City. One of the characteristics of the Kumamoto earthquake was that frequent small-scale aftershocks occurred subsequent to the main quake. If we were to include minor aftershocks that could not even be felt by people, then as many as 130,000 had occurred in the Kyushu region. Even those earthquakes with a magnitude of 3.5 or more have been recorded 339 times. Aftershocks are a common feature of earthquakes and occurred with extraordinary frequency in the case of the Kumamoto earthquake.

This study analyzes seismic activity using the Sentinel Application Platform (SNAP), which was developed under the ESA for the analysis of data obtained using InSAR. We used another software called SNAPHU distributed by SORSEFORGE to process the phase unwrapping approach data. In this study, the Sentinel satellite data from the C-band were compared with the ALOS satellite data from the L-band to check the performance of the C-band and L-band. We used data from the analysis of the L-band carried out by the Geographical Survey Institute regarding the Kumamoto earthquake (Geospatial Information Authority of Japan, 2016).

The analysis shows that the displacements caused by the Kumamoto earthquake range from −25 cm to +211 cm and occurred not only in areas surrounding the epicenter but also in wider areas in the Kumamoto Prefecture. Thus, phase unwrapping processing can show diastrophism in an intelligible way.

Generally, in order to guarantee the model verification, used data was split into a calibration area and a verification area. In this paper, however calibration and verification data were used as a same data. The reason why in-situ data of all Kumamoto area was small amount with only 9 points of GPS in around earthquake area. Since the relationship between calculated and in-situ displacement data is linear, the estimate model to derive displacement is as the follows:

$$D = -0.3114x + 10.599$$

where D is displacement, x should be calculated data with phase unwrapping. The general fit of the models was calculated as the root mean square deviation (RMSD), which describes the mean deviation of the estimated values, in the unit of measure used on the input data (cm). The RMSD on the test area was 6.8 with an R2 of 0.72.

Spatio-temporal nonlinear random effects modeling for COVID-19 case counts in Japan

Masao Ueki

Nagasaki University

Abstract:
Coronavirus disease 2019 (COVID-19) caused by the SARS-CoV-2 virus has spread seriously throughout the world. In Japan, the spread of infection still continues and the number of infected people is increasing. Statistical modeling of the number of COVID-19 cases may help understand the pattern of spread and predict future spread of the infection. In this study, we describe a spatio-temporal modeling approach by Paul and Held (2011, Stat Med 30:1118-36) to model COVID-19 case count data. The Paul-Held model allows to model the multivariate count time-series data in 47 prefectures of Japan simultaneously by incorporating spatial information between prefectures. It involves nonlinear random effects to capture the heterogeneity in model parameters associated with the prefectures. We apply the Paul-Held spatio-temporal nonlinear random effects model to COVID-19 case counts in Japan with the negative binomial distribution, and describe the results of the application.

(This research was partially funded by Nagasaki University “Doctoral Program for World-leading Innovative and Smart Education” for Global Health, KENKYU SHIDO KEIHI.)

Non-linear variable selection with ultra high-dimensionality

Yuta Umezu

Nagasaki University

Abstract:
In many area including statistics and machine learning, screening method is a kind of fundamental technique for handling large scale data, in which the number of variables is often larger than the sample size. In this talk, I explain some advances in screening method for detecting non-linear relationships between response and covariates in ultra high-dimensional scenario.

Geographic factor analysis of hotel plan prices using web booking data

Yu Ichifuji

Nagasaki University

Abstract:
In the era of open data, an increasing number of micro geo-data are becoming available. While remotely sensed images, people flow data, pricing data at gas stations, and many other micro geo-data will be useful for regional analysis, they tend to be noisy and have large samples. Computationally efficient statistical approach for noisy observations is increasingly important for modeling these data. In this research, our target is the tourism of Nagasaki Prefecture. The purpose of this study is to clarify the relationship between geographic factors in the occupancy rate and price of each hotel in Nagasaki Prefecture before the tourism industry was damaged by the COVID-19. This is an essential approach to understand how much recovery has been possible and how the situation has changed after the effects of the COVID-19 have subsided. Specifically, we extended the spatial econometric model based on occupancy rates estimated using data that can be collected via API from hotel booking websites to visualize the impact of each hotel. The data records occupancy rates and average unit price in 59 hotels in 814 days between July 1, 2017, and September 24, 2019. The results showed that each location has its own characteristics, such as areas that are greatly affected by nearby prices and occupancy rates, and areas where occupancy rates increase when prices decrease.

Development of Immersive VR Learning Equipment for Peace Education with Tangible User Interface

Norio Setozaki

Nagasaki University

Abstract:

More than 75 years have passed since the end of World War II, and the number of A-bomb survivors is aging and decreasing. Therefore, there is a need to consider how to provide peace education for the children who will lead the next generation. In this study, we developed an immersive VR learning equipment for peace education with Tangible User Interface. It is expected to increase interest in the damage caused by the atomic bombings and to reduce psychological distance through exploratory learning in a virtual environment created by the learners themselves.

Using maps and photographs of the city before the atomic bombing as reference, learners place models of houses on a table and recreate the streets of Nagasaki before the bombing. The streetscape reproduced by the models of houses is constructed in real time in a virtual environment. By wearing a Head Mounted Display (HMD), learners can enter the "model world" that they have created. Then, by using the controllers held in both hands, they can switch to the "pre-bombing" version, which shows the details of the streets and the sense of life at that time. Furthermore, using the controllers in both hands, learners can also switch to the "after the bombing" version.

Degradation Scheme by Geometrically-Transformed Image against Average Attack

Makoto Fujimura

Nagasaki University

Abstract:

Image contents are published on web via internet. It becomes easy to access and enjoy many multimedia contents. But problem of illegal copy of multimedia contents rise up concurrently. And some watermarking techniques are proposed for copyright protection and authentication of digital media. Digital watermarking is to embed auxiliary information into a host digital signal by imposing imperceptible signal changes. However, illegal user attacks the watermarking of copyright protection system to eliminate embedded watermark. Attacks on digital watermarking system are several types which are removal attacks, geometrical attack, cryptographic attack and protocol attack.

In recently years, multiuser collusion attacks are appeared and spread for cost-effective attacking method for removing watermarking. In case of traditional watermarking techniques, different identification information is embedded into each distributing image data of same content. This is reason of collusion attack, which is averaging attack especially, is possible to eliminate identification information embedded by watermarking techniques. Average attack is easy and efficiency method for eliminating embedding identification information. Average of same digital image contents can eliminate only embedding information, because different information is inserted at same location of same image content.

If distributed digital image data of same contents are different each other, average attack invokes image data quality degradation. It is possible to create different image data of same content using seam carving technique. Seam carving was proposed effective content-aware resizing of image without geometric constraints. Seam carving process is consisted of two parts. At first, seam lines of image are searched using evaluation function of energy according to content-aware. Next, seam lines are eliminated from original image and rest of regions are connected by translation. Created resizing image by seam carving keep image quality subjectively and is different from original image in each pixel sampling location. Therefore, we applied seam carving technique to the degradation scheme against average attack.

This degradation scheme is based on canceling out linear condition of digital filter for breaking down average attack.

Real-time signal and image processing with FPGA computing

Yuichiro Shibata

Nagasaki University

Abstract:
From a computer architecture perspective, one of technical challenges for machine learning and data science applications is handling of massive amount of data. Especially for cyber-physical systems such as autonomous robots, high-throughput input data produced by sensor devices must be processed within a low latency under a limited power constraint. In traditional computing approaches, the input data are firstly stored in memory with direct memory access (DMA) transfer, and then processed by a processor. Since the processor must wait for the DMA transfer to finish before starting the data processing, this imposes an inherent latency. In addition, the frequent memory access for data buffering tends to consume a large portion of the power budget. Field programmable gate array (FPGA) computing is an alternative approach. In this approach, by making the best use of re-programmability of FPGAs, deeply pipelined custom arithmetic hardware structure is configured for each application, enabling stream processing of input data. Since the input data can be directly fed into the pipeline without storing the memory, low latency is easily achieved. Moreover, data acquisition and processing can be overlapped in a pipelined manner, achieving high-throughput data processing. However, all the traditional computing algorithms do not fit with this FPGA computing concept. Thus, proper selection and modification of algorithms are often crucial issues as well as sophisticated design of hardware structure. In this talk, several implementation examples of signal and image processing applications are introduced to reveal what factors of algorithms affect the efficiency. Some FPGA-specific techniques to optimize the relationship between algorithms and architectures are also discussed.

"Low-rank and sparse" everywhere: what if plugged in deep learning?

Tomoya Sakai

Nagasaki University

Abstract:
Observation of structural events or spatially structured patterns likely derives a sequence or collection of data with recurring and redundant features. Unusual/abnormal events of interest, in contrast, cause sparse outlying features. The mixture of such redundant and sparse features in a high dimensional space can be represented by a low-rank and sparse (L+S) model in many cases. We will revisit the fundamentals and applications of L+S modeling. The L+S model applies well to a variety of time-series data such as surveillance video (background+foreground), optical-flow sequences (egomotion+object motion), lung sounds (breathing+wheezing), digital angiography (breathing deformation+blood flow), etc. We will further discuss some key benefits of a novel approach to introduce L+S prior knowledge into deep learning in those applications.