Frequency Domain Predictive Modelling with Aggregated Data

Avradeep Bhowmik
UT Austin
avradeep.1@utexas.edu

Joydeep Ghosh
UT Austin
ghosh@ece.utexas.edu

Oluwasanmi Koyejo
UIUC
sanmi@illinois.edu

Abstract

Existing work in spatio-temporal data analysis invariably assumes data available as individual measurements with localised estimates. However, for many applications like econometrics, financial forecasting and climate science, data is often obtained as aggregates. Data aggregation presents severe mathematical challenges to learning and inference, and application of standard techniques is susceptible to ecological fallacy. In this manuscript we investigate the problem of predictive linear modelling in the scenario where data is aggregated in a non-uniform manner across targets and features. We introduce a novel formulation of the problem in the frequency domain, and develop algorithmic techniques that exploit the duality properties of Fourier analysis to bypass the inherent structural challenges of this setting. We provide theoretical guarantees for generalisation error for our estimation procedure and extend our analysis to capture approximation effects arising from aliasing. Finally, we perform empirical evaluation to demonstrate the efficacy of our algorithmic aproach in predictive modelling on synthetic data, and on three real datasets from agricultural studies, ecological surveys and climate science.

1 Introduction

Analysis of spatio-temporally correlated data is an important and ever present problem in diverse and wide-ranging fields including econometrics [1], climate science [2, 3], financial forecasting [4] and Internet of Things (IoTs) [5, 6]. Nearly all existing modelling techniques in literature assume access to datasets with individual level samples for each time and/or location index. However, in many real life cases[7, 2, 1], for various reasons including measurement fidelity, robustness to random noise, cost of data collection, privacy preservation, scalability, etc., data is often collected and/or publicly reported as *aggregates* or *time averages*, collected over specific intervals and released periodically, e.g., data released by the Bureau of Labour Statistics [8] and US Department of Commerce [9], or by the General Social Survey [10] are often in this form.

The central question addressed in this paper is whether one can provably learn individual level models given only aggregated spatio-temporal data– a challenging and relatively unexplored form of semi-supervision, which requires novel techniques and significant algorithmic innovation on the part of data analysts to perform modeling and inference. As a first work (to the best of our knowledge) on predictive modelling with spatio-temporally aggregated data, we tackle the problem in the context of predictive linear modelling where real valued targets are regressed on multivariate features via a vector parameter.

Even for this relatively simple setup, naive application of standard modelling techniques to aggregated data often fails due to ecological fallacy [11, 12, 13] wherein inferences drawn at the group level differ significantly from the ground truth at individual level. Learning is especially difficult if aggregation periods are not uniform or aligned across features and targets. For example, an econometric model may want to use as features metrics like GDP growth rate (reported quarterly), unemployment rate and inflation rate (reported monthly), interest rate and balance of trade (reported daily) and ratio of government debt to GDP (reported yearly) to predict, say, stock market indices and currency exchange rates (reported daily) [9, 8].

In such a scenario, it is extremely challenging even to formulate a cogent mathematical representation

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

that captures the relationship among the available misaligned aggregates. On the other hand, effective reconstruction of data at the individual-level is very difficult because aggregation fundamentally obfuscates local information.

1.1 Contributions

In this paper, we demonstrate that by formulating the problem in the frequency domain, selected global properties of individual components of the model can be separately estimated with high fidelity even from aggregated data, which can then used for learning and inference without being affected by local-level information obfuscation caused by aggregation– all of this without any explicit data reconstruction. Our specific contributions are summarised below-

1. To our knowledge, we are the first to investigate the problem of predictive modelling from aggregated spatio-temporal data. We introduce a novel framework and new algorithmic mechanisms for learning from aggregated spatio-temporal data that leverages structural properties of frequency domain analysis techniques to perform predictive modelling with minimal data reconstruction.
2. We provide theoretical guarantees for our framework, and establish that under mild regularity conditions, the parameter vector learned from aggregated data suffers a generalisation error that is provably close to the optimal that can be obtained from any linear model in the non-aggregated setting, that is, with individual level samples
3. We extend our analysis to derive guarantees for our algorithm to capture real world approximation effects caused by aliasing and randomness in the data generation procedure, and show that our methods can still learn a parameter that closely matches the optimal generalisation error.

We empirically evaluate the efficacy of our methods on both synthetic data and three real datasets involving applications in ecological surveys, agricultural studies and climate science.

1.2 Related Work

There is a vast range of work on spatio-temporal data analysis [2, 14, 15] but very little existing literature applies to the aggregated case. The closest that come to our setup are interpolation techniques like Kriging [16, 17], which also typically assume that data is sampled at localised discrete positions on a grid, rather than as aggregates. Among frequency domain techniques, the closest line of work is spectral regression [18, 19, 20] which has been previously used in econometrics and financial modeling. However, existing work only deals with non-aggregated data in the discrete domain, and in particular, we have not come across an estimation framework nor analysis techniques, nor any guarantees for generalisation error as introduced in this manuscript.

There is limited existing literature in general for aggregated data of any kind. In the classification setting, learning from label proportions or LLP [21, 22] estimates classifiers from proportions of discrete valued labels in groups of labeled targets. Regression involving aggregated data was recently studied in [23] and [24] which considered the cases where data was aggregated into histograms and moments respectively. While the aforementioned pieces of work involved uniform aggregation of uncorrelated data, our setup involves data that is non-uniformly aggregated and spatio-temporally correlated. Moreover, our methods deal with aggregating continuous signals while the existing work outlined above involves aggregation of discrete values.

Note that while our work involves spatio-temporal data, the goal is nevertheless a general framework for predictive modelling rather than forecasting– in fact, our methods can be used even outside spatio-temporal applications, e.g. in any domain wherein sampled measurements can be represented as tensors where a sense of ordering or structural chronology exists along each mode (for example, clinical measurements).

2 Preliminaries

Before we go into the specifics of the estimation process, we recall some fundamental results and quantities from Fourier analysis (see [25, 26]). A signal or stochastic process $z(t)$ defined on $t \in \mathbb{R}$ is **centred** and **weakly stationary with finite variance** if:

1. the process is centred, $E[z(t)] = 0$ for all t
2. for any t, t', we have $E[z(t)z(t')] = \rho_z(\|t - t'\|)$ for a non-negative real valued auto-correlation function $\rho_z(\cdot) : \mathbb{R}^+ \mapsto \mathbb{R}^+$
3. at every point, the process has finite variance, $E[z(t)^2] = \rho(0) < +\infty$

Given a continuous signal $z(t)$, the **Fourier Transform** of the signal with respect to a particular frequency $\omega \in \mathbb{R}$ is given by

$$Z(\omega) = \int_{\mathbb{R}} z(t) e^{-\iota\omega t} dt \tag{1}$$

For a signal $z(t)$, we use both $Z(\omega)$ and $\mathcal{F}z(\omega)$ to denote its Fourier transform.

We can similarly define the ***T*-restricted Finite Fourier Transform** $Z_T(\omega)$ for the signal $z(t)$ as

$$Z_T(\omega) = \int_{-T}^{T} z(t)e^{-\iota\omega t}dt \tag{2}$$

The **Power Spectral Density** $P_Z(\omega)$ of a signal $z(t)$ with respect to a particular frequency $\omega \in \mathbb{R}$ is given by

$$P_Z(\omega) = \lim_{T\uparrow\infty} \frac{1}{T} E\left[\| \int_{-T}^{T} z(t)e^{-\iota\omega t}dt \|^2 \right] \tag{3}$$

Let $z(t)$ be a weakly stationary process with autocovariance function $\rho_z(\tau) = E[z(t)z(t+\tau)]$. Let $\rho_z(0) = E[z(t)^2] < \infty$ be the variance of the process. We simply state the following well known results without proof [26]:

1. (Wiener-Khinchin Theorem)The power spectral density of a stationary process $z(t)$ is the Fourier Transform of its autocovariance function

$$P_Z(\omega) = \int_{-\infty}^{\infty} \rho_z(\tau)e^{-\iota\omega(\tau)}d\tau \tag{4}$$

2. (Corrollary of above) For a stationary process, the integral of the power spectral density gives the instantaneous variance

$$\int_{-\infty}^{\infty} P_Z(\omega)d\omega = \rho_z(0) = E[z(t)^2] \tag{5}$$

These results are examples of the well known duality properties of Fourier analysis, where global properties in the time domain are related to local properties in the frequency domain and vice versa. We shall use these properties extensively in our work.

Throughout this manuscript, we assume that the power spectral density (and correspondingly, the autocovariance function) for every signal of interest exists finitely, and decays rapidly with lag for all processes involved. In particular, we assume that $\rho_{(\cdot)}(\cdot)$ is a Schwartz function [27], that is $\rho(\cdot)$ and all its derivatives decay at least as fast as any inverse polynomial. Therefore, most of the power for our signals will be concentrated around $\omega = 0$. An extended discussion on this is presented in section III in the **supplement**.

3 Problem Setup

In the interest of simplicity we delineate our setup for temporally aggregated data, where features $\mathbf{x}(t)$ and targets $y(t)$ are time series signals or processes. Discussion on higher dimensional aggregation frameworks are deferred to section 5.

Consider the task of predictive linear modelling, where real valued targets $y(t) \in \mathbb{R}$ are regressed on multivariate feature vectors $\mathbf{x}(t) \in \mathbb{R}^d$ via a parameter vector $\boldsymbol{\beta}^* \in \mathbb{R}^d$ in a linear model

$$y(t) = \mathbf{x}(t)^\top \boldsymbol{\beta}^* + \epsilon(t) \tag{6}$$

where $\epsilon(t)$ is a random noise process. For the rest of our manuscript, we make the assumption that all our signals of interest $\mathbf{x}, y, \epsilon$ are centered and weakly stationary with finite variance.

Stationarity is a standard assumption in time series analysis and very common in many real life applications (see [28, 29, 30]), and techniques like filtering out trend lines and differencing are often applied to the data to ensure stationarity before analysis [31]. Note that we do not assume any specific functional form for the generative processes (Gaussian, etc.) for the signals studied in this manuscript.

Loss Function and Parameter Estimation

Standard statistical learning approaches estimate the optimal linear model given the data by minimising an appropriate loss function over the vector parameter $\boldsymbol{\beta}$. Define the residue process at any particular $\boldsymbol{\beta}$ as

$$\varepsilon_\beta(t) = \mathbf{x}(t)^\top \boldsymbol{\beta} - y(t)$$

One potential option for a loss function might have been the total energy of the residue process $\int_{\mathbb{R}} |\varepsilon_\beta(t)|^2 dt$

However, the total energy in the noise process is often not finite [25, 32], hence for weakly stationary processes, a better loss function to use is the variance of the noise process at time t, that is,

$$\mathcal{L}(\boldsymbol{\beta}) = E[|\varepsilon_\beta(t)|^2] = E[|\mathbf{x}(t)^\top \boldsymbol{\beta} - y(t)|^2]$$

By assumption our signals are weakly stationary, therefore the variance does not depend on t. Therefore, the "optimal" linear model parameter is given by

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} E[|\mathbf{x}(t)^\top \boldsymbol{\beta} - y(t)|^2] \tag{7}$$

Given access to the detailed, full-resolution dataset, the typical strategy for solving the estimation problem (7) is to replace the expectation by a sum over individual datapoints. This finite sum converges to the expectation given enough datapoints under certain conditions, for example, if the noise process is ergodic [33]. However, the story becomes more complicated if the data is available in aggregated form.

3.1 Data Aggregation in Time Series

Instead of the individual targets $y(t)$ at time t, we are given aggregates sampled with period T, which are of the form

$$\overline{y}[k] = \frac{1}{T} \int_{(k-1)T/2}^{kT/2} y(\tau) d\tau \tag{8}$$

for $k \in \mathbb{Z} = \{\cdots -1, 0, 1, \cdots\}$.

Features can also be aggregated in a more complicated manner with different periodicities, that is, each coordinate $\{\overline{x}_i(t) : \ i = 1, 2, \cdots d\}$ of the features $\mathbf{x}(t)$ can be aggregated periodically with period T_i as

$$\overline{x}_i[l] = \frac{1}{T_i} \int_{(l-1)T_i/2}^{lT_i/2} x_i(\tau) d\tau \tag{9}$$

Therefore, instead of the continuous time data $\{(\mathbf{x}(t), y(t)) : t \in \mathbb{R}\}$ specified across t, we are given access to discrete aggregates $\{\overline{y}[k] : k \in \mathbb{Z}\}$ and sets of aggregates $\{\{\overline{x}_i[l] : i = 1, 2, \cdots d\} : l \in \mathbb{Z}\}$.

4 Frequency Domain Parameter Estimation from Aggregated Data

We show that an approximately equivalent frequency domain formulation of the problem allows us to sidestep the challenges inherent in a data aggregation setup without explicit reconstruction. Since local time-domain properties are captured by global frequency domain properties, a frequency domain analysis allows us to individually extract high fidelity estimates of selected global properties of all the quantities involved to then use for inference and predictive modelling.

4.1 Frequency Domain Representation of Aggregated Time-Series Data

The first key insight that enables us to work with aggregated time series data is the fact that aggregation in the time domain corresponds to convolution and subsampling in the frequency domain.

Recall that in our setup, continuous signals of the form $z(t)$ get aggregated into samples of the form-

$$\overline{z}[k] = \frac{1}{T} \int_{(k-1)T/2}^{kT/2} z(\tau) d\tau \tag{10}$$

There are two steps here. First, the continuous process $z(t)$ is aggregated into the sliding-window averaged continuous process $\overline{z}(t)$ as

$$\overline{z}(t) = \frac{1}{T} \int_{t-T/2}^{t+T/2} z(\tau) d\tau \tag{11}$$

This is equivalent to a convolution operation $\overline{z}(t) = z(t) * u(t)$ with the square wave function $u(t) = \frac{1}{T}\mathbb{I}\{t \in (-T/2, T/2)\}$, where $\mathbb{I}\{\cdot\}$ is the indicator function. In the frequency domain, this is equivalent to multiplying with a sinc function $U_T(\omega) = \frac{sin(\omega T/2)}{\omega T/2}$.

The final observation sequence $\{z[k] : k \in \mathbb{Z}\}$ is obtained by sub-sampling at periodicity T the aggregated time series $\overline{z}(t)$; in the frequency domain this becomes a $\frac{2\pi k}{T}$-periodicity sub-sampling operation, via a convolution with a delta train or a Dirac comb $\frac{1}{T}\sum_{k\in\mathbb{Z}} \delta(\omega - \frac{2\pi k}{T})$.

Therefore, putting it all together, we can write our observation signal in the frequency domain as

$$\bar{Z}(\omega) = \frac{1}{T} \sum_{k\in\mathbb{Z}} Z(\omega - \frac{2\pi k}{T}) U_T(\omega - \frac{2\pi k}{T}) \tag{12}$$

$$= \frac{1}{T} Z(\omega) U_T(\omega) + \Delta_z(\omega|T) \tag{13}$$

where $\Delta_z(\omega|T) = \frac{1}{T}\sum_{k\in\mathbb{Z}\setminus\{0\}} Z(\omega - \frac{2\pi k}{T}) U_T(\omega - \frac{2\pi k}{T})$ is the error due to aggregation and **aliasing**.

For succinctness of notation, we assume identical rates for aggregation and subsampling. Estimation is identical in the case where aggregation time period and reporting frequency are different for targets and features (e.g. in case of overlapping aggregation or sliding windows), but the analysis requires some additional book-keeping - a brief discussion is included in section 5.

4.2 Formulation and Estimation Algorithm

We now proceed to formulate our parameter estimation framework in the frequency domain. First, we note that the Fourier[1] Transform $z \leftrightarrow \mathcal{F}z$ is a linear operation, therefore the linear relationship that holds in the time domain must also hold in the frequency domain. That is for any signal $\mathbf{x}(t), y(t)$ with noise $\epsilon(t)$, and for any $\boldsymbol{\beta}$, we have

$$y(t) = \mathbf{x}(t)^\top \boldsymbol{\beta} + \epsilon(t) \iff Y(\omega) = \mathbf{X}(\omega)^\top \boldsymbol{\beta} + \mathcal{E}(\omega)$$

Therefore, it stands to reason that if we have good estimates for $Y(\omega), \mathbf{X}(\omega)$ for specific values of ω, parameter estimation should be able to proceed in the frequency domain.

However, the preceding section makes it clear that unless our signals are band-limited, estimates for $\mathbf{X}(\omega)$

[1] as well as the Finite Fourier Transform

Algorithm 1 Fourier-domain Estimation from Aggregated Data

Input: $\overline{x}, \overline{y}, \omega_0, D, T_0$

Sample D frequencies uniformly in $(-\omega_0, \omega_0)$ to get

$$\Omega = \{\omega_1, \omega_2, \cdots \omega_D : \omega_i \in (-\omega_0, \omega_0)\}$$

for each $\omega \in \Omega$, and $i \in \{1, 2, \cdots d\}$ **do**

compute the T_0-limited finite Fourier transforms

$$\overline{X}_{i,T_0}(\omega) = \mathcal{F}_{T_0}\overline{x}_i(\omega),\ \overline{Y}_{T_0}(\omega) = \mathcal{F}_{T_0}\overline{y}(\omega)$$

reconstruct non-aggregated Fourier Transforms

$$\widehat{X}_{i,T_0}(\omega) = \frac{\overline{X}_{i,T_0}(\omega)}{U_{T_i}(\omega)},\ \widehat{Y}_{T_0}(\omega) = \frac{\overline{Y}_{T_0}(\omega)}{U_T(\omega)}$$

end for

Estimate the parameter as

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{|\Omega|} \sum_{\omega \in \Omega} \|\widehat{\mathbf{X}}_{T_0}(\omega)^\top \boldsymbol{\beta} - \widehat{Y}_{T_0}(\omega)\|^2$$

return $\widehat{\boldsymbol{\beta}}$

and $Y(\omega)$ will be affected by aliasing. Since in the real world we can only work with finite time signals, our signals will never be band-limited because they are time-limited.

Nevertheless, if we assume that the power spectral density for the original signal decays rapidly with ω and, for some ω_0, almost vanishes beyond $|\omega| > \omega_0$. Then, it is easy to see that the effect of aliasing from the sampling process will be minimum for all our signals around $\omega = 0$. Therefore, it makes sense to use only high fidelity estimates of $Y(\omega), \mathbf{X}(\omega)$ for estimation, by restricting ourselves to values of $\omega \in (-\omega_0, \omega_0)$.By doing so, we also bypass any necessity for reconstruction of the original values of our signals in the time domain.

These ideas are the crux of the intuition for our framework and algorithmic treatment of the problem. Section I in the supplement contains an extended expository discussion that motivates and outlines the steps involved in translating these intuitive ideas to specific algorithmic strategy in mathematical terms.

By formulating the estimation problem in the frequency domain in a way that exactly exploits these intuitive ideas, we can derive our first main result which shows that under our assumptions, frequency domain parameter estimation leads to generalisation error that is close to the optimal.

Theorem 1. *Let $\boldsymbol{\beta}^*$ be the optimal parameter as in equation 7. Denote the parameter estimated from the T_0-restricted Fourier Transforms as*

$$\widehat{\boldsymbol{\beta}} = arg\min_{\boldsymbol{\beta}} \sum_{\omega \in \Omega} E\left[\|\mathbf{X}_{T_0}(\omega)^\top \boldsymbol{\beta} - Y_{T_0}(\omega)\|^2\right] \quad (14)$$

Then, for every small $\xi_1, \xi_2 > 0$, there exist correspondingly T_0, ω_0, D such that for the set $\Omega = \{-\omega_0 < \omega_i < \omega_0 : i = 1, 2, \cdots |\Omega|\}$ with $|\Omega| = D$ sampled uniformly between $(-\omega_0, \omega_0)$, we have

$$E\left[|\mathbf{x}(t)^\top \widehat{\boldsymbol{\beta}} - y(t)|^2\right] < (1+\xi_1)\left(E\left[|\mathbf{x}(t)^\top \boldsymbol{\beta}^* - y(t)|^2\right]\right) + (1+\xi_1)\xi_2$$

with probability at least $1 - e^{-O(D^2\xi_2^2)}$

In essence, this result shows that given a long enough signal, with enough granularity in sampled frequencies, the estimated parameter $\widehat{\boldsymbol{\beta}}$ leads to a generalisation error that is arbitrarily close to the optimal generalisation error obtained by $\boldsymbol{\beta}^*$. Because of the multiple tunable parameters in our formulation, it allows for enough trade-offs that our algorithm can be applied to a wide range of applications (see for example, [34, 35, 36, 37]), and Theorem 1 can be used as a generic template to derive more precise and bespoke guarantees for each such case. The exact guarantees obtained will depend on the specifics of the application and the data setup– we provide a concrete example of a particular class of common cases in the subsequent section.

4.3 Aliasing and Approximation Effects

In real life cases, we have to deal with approximation effects arising from aliasing and randomness of the data that affect our algorithm and analysis procedure, especially in computing our objective function. However, we can show that in most cases the objective function in our estimator as defined in equation (14) can be closely approximated with mild regularity assumptions.

For instance, suppose we have data collected independently from N locations with corresponding T_0-restricted Fourier Transforms $\{(\mathbf{X}^j_{T_0}(\omega), Y^j_{T_0}(\omega)) : j = 1, 2, \cdots N\}$ (for example, these can be economic metrics from different states or counties, or meteorological measurements at different points in the atmosphere). We assume that the individual processes at each location is strictly sub-Gaussian [38, 39]. We also assume that the power spectral density of all processes involved is finite for every $\omega \in (-\omega_0, \omega_0)$, and decays rapidly at a sub-Gaussian rate $e^{-O(\omega-\omega_0)^2}$ beyond $|\omega| > \omega_0$.

Then, the following result holds which shows that even for the case where the targets and features are aggregated at different rates, we can still estimate a parameter that leads to a generalisation error that is close to the optimal linear modelling error.

Theorem 2. *Let T_i be the sampling/aggregation period for the i^{th} coordinate $x_i(t)$ and T_y be the corresponding period for the target $y(t)$. Let $\omega_s = \frac{2\pi}{T_s}$ with $T_s =$*

$\max\{T_y, T_1, T_2, \cdots T_d\}$. *Denote the parameter obtained by our estimator from N data sources as*

$$\widehat{\boldsymbol{\beta}} = arg\min_{\boldsymbol{\beta}} \sum_{j\in[N]} \sum_{\omega\in\Omega} \|\widehat{X}^j_{T_0}(\omega)^\top \boldsymbol{\beta} - \widehat{Y}^j_{T_0}(\omega)\|^2$$

Then, for every small $\xi_1, \xi_2, \xi_3 > 0$, there exist correspondingly T_0, ω_0, D such that for the set $\Omega = \{-\omega_0 < \omega_i < \omega_0 : i = 1, 2, \cdots |\Omega|\}$ with $|\Omega| = D$ sampled uniformly between $(-\omega_0, \omega_0)$, we have, if the aggregation rate is high enough $\omega_s > 2\omega_0$,

$$E\left[|\mathbf{x}(t)^\top \widehat{\boldsymbol{\beta}} - y(t)|^2\right] < (1+\xi_1)\left(E\left[|\mathbf{x}(t)^\top \boldsymbol{\beta}^* - y(t)|^2\right]\right) + (1+\xi_1)(\xi_2 + \xi_3 + e^{-O((\omega_s - 2\omega_0)^2)})$$

with probability at least $1 - e^{-O(D^2\xi_2^2)} - e^{-O(N^2\xi_3^2)}$

Note that our estimation procedure requires no explicit reconstruction of the original time domain data, which would require spectral information about the signal over the entire spectrum, much of which is severely affected by aliasing effects. In contrast, our methods only use information about the specific parts of the spectrum which are robust and least impacted by aliasing, and are thus more accurate snapshots of the signal.

When the sampling and aggregation periodicity is uniform across all coordinates, an interesting effect can be observed wherein uniform aliasing effects in features and targets essentially cancel each other out. This is because the aliasing error Δ_x for features are related linearly to the error Δ_y for targets via the same parameter. Therefore, parameter estimation can proceed without explicit reconstruction of $\widehat{X}_i(\omega), \widehat{Y}(\omega)$ as a standard linear regression albeit with a slightly different noise model. However, estimation can still be affected by aliasing in the noise in the signal, therefore, as our experiments on synthetic data shall show, it may preferable to perform estimation in the frequency domain nevertheless.

5 Discussion and Extensions

5.1. Multi-dimensional Aggregation:

So far our discussion has been limited to the case where d-dimensional feature vectors $\mathbf{x}$ and real valued targets y are obtained at (and aggregated along) points on a single dimension, i.e., time. We can extend our work very easily to the more general case, where features and targets are indexed by and averaged over points in the p-dimensional Euclidean space $\mathbb{R}^p$.

For example, in spatial climate models, we may use as features $\mathbf{x} \in \mathbb{R}^d$ and targets $y \in \mathbb{R}$ values of meteorological variables (CO_2 levels, temperature, etc.) at discrete points on the earth's surface, indexed by a 2-dimensional (latitude, longitude) vector (i.e., $p = 2$). But instead of $(\mathbf{x}, y)$ for every location, measurements may only be available aggregated averaged over regions on the earth's surface (e.g., averages over 1mi x 1mi spatial grids), which can then be used for learning climate models. Similarly, in 3-dimensional space, $p = 3$, measurements can be obtained aggregated over 3-d blocks. Note that the ambient dimension p is distinct from the dimensionality of the feature space d.

Suppose locations in $\mathbb{R}^p$ are indexed by points $\mathbf{v}$, and each such location is associated with its own d-dimensional feature vector $\mathbf{x}(\mathbf{v}) \in \mathbb{R}^d$ and real valued target $y(\mathbf{v}) \in \mathbb{R}$, which are regressed on each other via a vector parameter $\boldsymbol{\beta}^* \in \mathbb{R}^d$ as

$$y(\mathbf{v}) = \mathbf{x}(\mathbf{v})^\top \boldsymbol{\beta}^* + \epsilon(\mathbf{v}) \tag{15}$$

Each signal here is again a random zero-mean, weakly stationary noise process with finite variance. Observations for any signal[2] $\mathbf{z}(\mathbf{v})$ are again obtained as aggregates over periodically translated bounded connected set $A \subset \mathbb{R}^p$ as

$$z[\mathbf{k}] = \frac{1}{Vol(A)} \int_{\mathbf{v}\in A+\mathbf{k}} z(\mathbf{v}) d\mathbf{v}$$

Given a signal $z(\mathbf{v})$, for any "frequency" vector $\boldsymbol{\theta} = [\theta_1, \theta_2, \cdots \theta_p] \in \mathbb{R}^p$, the **Multidimensional Fourier Transform** is defined in a way very similar to the one-dimensional case [32, 40, 41]

$$Z(\boldsymbol{\theta}) = \int_{\mathbb{R}^p} z(\mathbf{v}) e^{-\iota\langle\boldsymbol{\theta},\mathbf{v}\rangle} d\mathbf{v} \tag{16}$$

where $\langle\cdot,\cdot\rangle$ represents the standard inner product.

All properties of Fourier Transforms required within the scope of this manuscript follow exactly as in the unidimensional case (see [40, 41]). For example, aggregation over regions defined by periodic translations of a set $A \subset \mathbb{R}^p$ becomes equivalent to multiplication in the frequency domain with the corresponding multidimensional Fourier Transform of the indicator function $g_A(\mathbf{v}) = \mathbb{I}(\mathbf{v} \in A)$. In particular, if A is the hypercube $A = \{\mathbf{v} : -a_i/2 \le v_i \le a_i/2\}$, then $\mathcal{F}g_A(\boldsymbol{\theta}) = \prod_{i=1}^p U_{a_i}(\theta_i)$, where $U_{(\cdot)}$ is the standard sinc function as in the unidimensional case.

The algorithm and results remain virtually identical with unidimensional quantities being replaced by their multidimensional equivalents. The only penalty that we pay is the number of sampled frequencies required, that is $|\Omega|$, which can in some cases scale exponentially with p. However, we note that in most real life cases p is very small (limited to at most $p = 4$ for spatio-temporal applications), hence this is not a severe impediment on the application on our methods.

[2] where $\mathbf{z}(\mathbf{v})$ is a stand-in for either $\mathbf{x}(\mathbf{v})$ or $y(\mathbf{v})$

5.2. Sliding Windows:

The estimation protocol in this case remains unchanged, but the analysis involves a little extra book-keeping. Note that a sliding window basically means that the aggregation periodicity and sampling periodicity are different. Say T_a is the aggregation period, that is, the period over which averages are computed for the signal (as in equation (11)). Also let T_b be the sampling period, that is, the period with which the aggregated signal $\bar{z}(t)$ is sampled. Then, equation (12) can be rewritten as

$$\bar{Z}(\omega) = \frac{1}{T_b} \sum_{k \in \mathbb{Z}} Z(\omega - \frac{2\pi k}{T_b}) U_{T_a}(\omega - \frac{2\pi k}{T_b}) \qquad (17)$$

with a corresponding aliasing error term $\Delta_z(\omega | T_a; T_b) = \frac{1}{T_b} \sum_{k \in \mathbb{Z} \setminus \{0\}} Z(\omega - \frac{2\pi k}{T_b}) U_{T_a}(\omega - \frac{2\pi k}{T_b})$. Theorem 2 can then be extended to show that in general, if T_a is reasonably small relative to $\frac{2\pi}{\omega_0}$, the aliasing error is dominated by effects from T_b, the sampling period. However, if T_a becomes too large in comparison to $\frac{2\pi}{\omega_0}$, the sinc function $U_{T_a}(\omega)$ can become too sharp and peaky which may result in gaps in the spectrum covered by Ω (refer to the proofs in the supplement for more details). This is intuitive since larger aggregation windows lead to higher loss of information.

5.3. Aggregation with Weighted Smoothing:

The analysis in the paper has been presented in the context of a simple aggregation schema that uses a square wave as a smoothing function for averaging. To cater to alternative aggregation schemata, one just needs to replace the sinc function $U(\cdot)$ with the Fourier Transform of the specific aggregation scheme being used– e.g., for Gaussian smoothing, the relevant Fourier Transform will be another Gaussian, etc. Our results remain unchanged for Schwartz smoothing functions, which includes most of the commonly used smoothing functions. In particular, note that the Gaussian function is a Schwartz function, and so is any smoothing function over a finite support (square wave, triangular wave, etc.), therefore their Fourier Transforms are Schwartz functions as well.

6 Experiments

We empirically evaluate the efficacy of our methods on both synthetic data and three real datasets. In each case, we use an aggregated version of the individual-level dataset for learning model parameters using the techniques in this paper, and evaluate the results by computing the predictive error obtained by our parameter on the full non-aggregated dataset. Since this is a first work on this topic, we are unaware of any real algorithmic baselines. However, we do test our methods against two baselines- the "true" linear model which is learned with access to the full non-aggregated dataset, and a "time-domain" model that naively imputes individual-level measurements by substituting the corresponding average for the group.[3]

Our synthetic data experiments proceed as follows. We generate multivariate time series data as features $\mathbf{x}(t)$ and univariate time series data as targets $y(t)$ that obeys our assumptions in this manuscript. We then aggregated this data– first using uniform sampling frequency, and second using non-uniform sampling frequency with increasing average discrepancy in the periodicity across features and targets. The aggregated data is then used for learning our model, and the results are compared against a time domain method that imputes the individual aggregates with group level values.

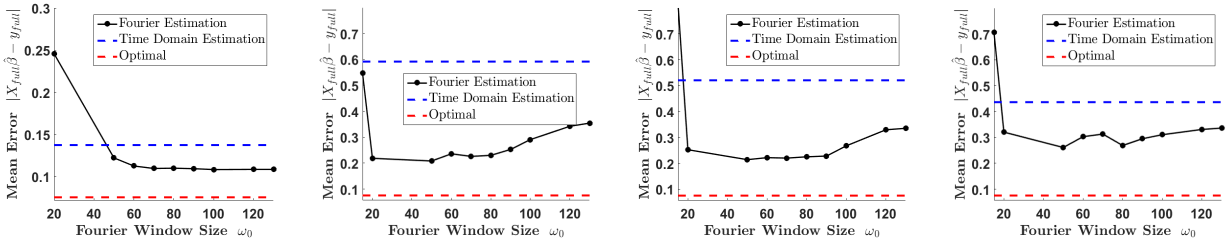
Plots for mean estimation error $|\mathbf{x}(t)\widehat{\boldsymbol{\beta}} - y(t)|$ with increasing Fourier Window ω_0 are shown for the uniform sampling period in figure 1a, and for non-uniform sampling period with increasing discrepancy in periodicity in figures 1b through 1d. In each of these cases, the results show that beyond a certain value of ω_0, frequency domain learning significantly outperforms naive time domain modeling. As described in section 4.3, Figure 1a shows that for uniform sampling frequency, time domain methods can be still used but our framework is nevertheless preferable because aliasing from error signal can affect estimation accuracy in the time domain. Moreover, as we describe in the manuscript, the performance of frequency domain estimation deteriorates if the value of ω_0 becomes too high because aliasing effects start distorting the results.

The first real spatio-temporal dataset involves an application from agricultural studies, wherein corn yield monitor data [42] from the Las Rosas agricultural plantation in Cordoba, Argentina is regressed against features including nitrogen levels, topographical properties, brightness value, etc. (see [43, 14] for further details on the dataset).

The second real dataset is the Forest Fires Dataset from the UCI Machine Learning Repository [44] which involves predictive modelling of burned acreage from forest fires in the northeast region of Portugal. by using as features meteorological and other data like relative humidity, ISI index, etc. (see [45] for more details on the dataset).

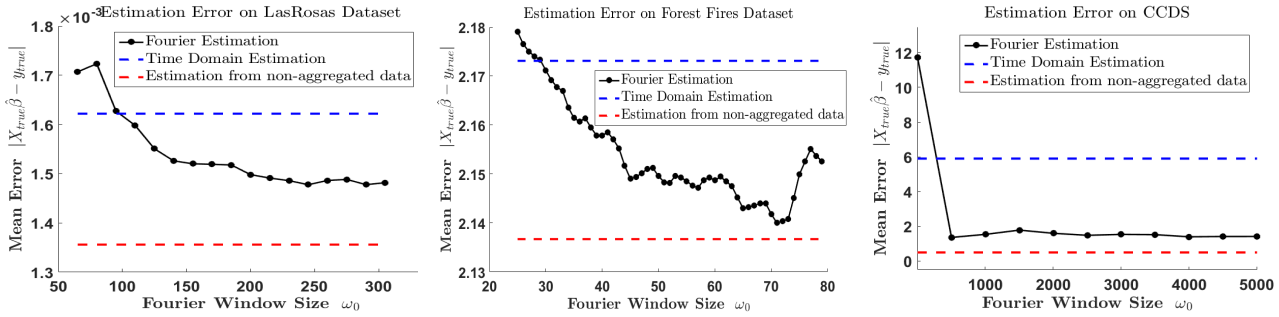
In both these datasets, the data points are stamped with latitude-longitude positional indices, which are used to topographically order each observation. The ordered data is then aggregated based on positional

[3] We also tried kriging for resampling i.e. reconstructing the non-aggregated data, then fitting a linear model on the resampled data. This approach performed poorly, hence we omit the results for clarity



(a) Uniform Sampling (b) Low Discrepancy (c) Medium Discrepancy (d) High Discrepancy

Figure 1: Results on Synthetic Data – Mean Estimation Error with increasing Fourier Window ω_0 for uniform aggregation (1a), and non-uniform aggregation with increasing discrepancy among aggregation periodicities (1b through 1d). Frequency domain parameter estimation outperforms naive application of time domain methods



(a) Las Rosas Dataset (b) Forest Fires Dataset (c) CCDS Dataset

Figure 2: Results on Forest Fires Dataset, Las Rosas Datasets show that frequency domain parameter estimation outperforms naive application of time domain methods and approaches the optimal for high enough ω_0. If ω_0 is too large, however, aliasing effects can lead to deteriorated performance as in Figure 2b

indices and used for learning a linear model.

In our final experiment, we test our techniques on the Comprehensive Climate Dataset (CCDS) which is an extensive collection of climate modeling variables for North America compiled from various sources including NASA, National Oceanic and Atmospheric Administration (NOAA), National Climate Data Center (NCDC), etc. (see [2, 3] for further details on the dataset). We use this dataset to model atmospheric vapour levels using various measurements, including carbon dioxide, methane, cloud cover, etc. and other extra-meteorological factors like rate of frost/rainy days, etc. over a grid that covers most of continental United States. This collection contains two datasets, one of which is aggregated and the other is observed at a much higher resolution. We use the aggregated dataset for learning $\widehat{\boldsymbol{\beta}}$ and test the predictive performance of our learned model on the higher resolution dataset.

Figures 2a, 2b and 2c show plots for mean estimation error $|\mathbf{x}(t)\widehat{\boldsymbol{\beta}} - y(t)|$ with increasing Fourier Window ω_0 for each of the three real datasets. Our results show that in all three datasets, for a large enough ω_0 our method significantly outperforms the corresponding time domain technique, and starts coming close to the performance of the optimal estimator.

7 Conclusion

In this manuscript we investigated the problem of predictive modelling of linear models involving correllated spatio-temporal data when the data is available only in aggregated form rather than as individual-level measurements with localised estimates. In particular, we analysed the scenario where aggregation is non-uniform across targets and different coordinates of the features, leading to significant challenges in cogent mathematical representation of any relationship among available feature and target aggregates. We showed that by formulating the problem in the frequency domain and exploiting duality properties of Fourier analysis, many of the inherent structural challenges of this setting can be bypassed. We introduced a novel framework and new algorithmic techniques to perform frequency domain estimation and inference for this setup and provided both theoretical guarantees and empirical validation of our methods. Future work will investigate extension of this paradigm to non-linear modelling, and estimation under alternative assumptions on data generation.

Acknowledgements

Authors acknowledge support from Verizon and NSF grant IIS 1421729.

References

[1] James EH Davidson, David F Hendry, Frank Srba, and Stephen Yeo. Econometric modelling of the aggregate time-series relationship between consumers' expenditure and income in the united kingdom. *The Economic Journal*, pages 661–692, 1978.

[2] Aurelie C Lozano, Hongfei Li, Alexandru Niculescu-Mizil, Yan Liu, Claudia Perlich, Jonathan Hosking, and Naoki Abe. Spatial-temporal causal modeling for climate change attribution. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 587–596. ACM, 2009.

[3] Yan Liu, Alexandru Niculescu-Mizil, Aurelie C Lozano, and Yong Lu. Learning temporal causal graphs for relational time-series analysis. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 687–694, 2010.

[4] Stephen J Taylor. Modelling financial time series. 2007.

[5] Li Da Xu, Wu He, and Shancang Li. Internet of things in industries: A survey. *IEEE Transactions on Industrial Informatics*, 10(4):2233–2243, 2014.

[6] Shancang Li, Li Da Xu, and Xinheng Wang. Compressed sensing signal and data acquisition in wireless sensor networks and internet of things. *IEEE Transactions on Industrial Informatics*, 9(4):2177–2186, 2013.

[7] Jenna Burrell, Tim Brooke, and Richard Beckwith. Vineyard computing: Sensor networks in agricultural production. *IEEE Pervasive computing*, 3(1):38–45, 2004.

[8] *Bureau of Labour Statistics, US Department of Labour*. http://www.bls.gov/.

[9] *Bureau of Economic Analysis, US Department of Commerce*. http://www.bea.gov/.

[10] *General Social Survey, NORC*. http://www3.norc.org/GSS+Website/.

[11] William S Robinson. Ecological correlations and the behavior of individuals. *International journal of epidemiology*, 38(2):337–341, 2009.

[12] David A Freedman, Stephen P Klein, Jerome Sacks, Charles A Smyth, and Charles G Everett. Ecological regression and voting rights. *Evaluation Review*, 15(6):673–711, 1991.

[13] Leo A Goodman. Ecological regressions and behavior of individuals. *American Sociological Review*, 1953.

[14] Dayton M Lambert, James Lowenberg-Deboer, and Rodolfo Bongiovanni. A comparison of four spatial regression models for yield monitor data: A case study from argentina. *Precision Agriculture*, 5(6):579–600, 2004.

[15] Joyce C Ho, Yubin Park, Carlos Carvalho, and Joydeep Ghosh. Dynacare: Dynamic cardiac arrest risk estimation. In *AISTATS*, pages 333–341, 2013.

[16] Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 2012.

[17] Margaret A Oliver and Richard Webster. Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information System*, 4(3):313–332, 1990.

[18] Deng Cai, Xiaofei He, and Jiawei Han. Spectral regression for efficient regularized subspace learning. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. IEEE, 2007.

[19] Peter CB Phillips et al. *Spectral regression for cointegrated time series*. Cowles Foundation for Research in Economics at Yales University, 1988.

[20] Dean Corbae, Sam Ouliaris, and Peter CB Phillips. Band spectral regression with trending data. *Econometrica*, 70(3):1067–1109, 2002.

[21] Novi Quadrianto, Alex J Smola, Tiberio S Caetano, and Quoc V Le. Estimating labels from label proportions. *The Journal of Machine Learning Research*, 10:2349–2374, 2009.

[22] Giorgio Patrini, Richard Nock, Tiberio Caetano, and Paul Rivera. (almost) no label no cry. In *Advances in Neural Information Processing Systems*, pages 190–198, 2014.

[23] Avradeep Bhowmik, Joydeep Ghosh, and Oluwasanmi Koyejo. Generalized Linear Models for Aggregated Data. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 93–101, 2015.

[24] Avradeep Bhowmik, Joydeep Ghosh, and Oluwasanmi Koyejo. Sparse parameter recovery from aggregated data. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1090–1099, 2016.

[25] Lambert H Koopmans. *The spectral analysis of time series*. Academic press, 1995.

[26] Loukas Grafakos. *Classical and modern Fourier analysis*. Prentice Hall, 2004.

[27] T TerzioĞglu. On schwartz spaces. *Mathematische Annalen*, 182(3):236–242, 1969.

[28] Clive William John Granger and Paul Newbold. *Forecasting economic time series*. Academic Press, 2014.

[29] Kacha Dzhaparidze. *Parameter estimation and hypothesis testing in spectral analysis of stationary time series*. Springer Science & Business Media, 2012.

[30] Edgar L Feige and Douglas K Pearce. The causality relationship between money and income: A time series approach. *Quarterly Journal of Business and Economics*, 13(4):183, 1974.

[31] Michael Hibon and Spyros Makridakis. Arma models and the box–jenkins methodology. 1997.

[32] Arun K Tangirala. *Principles of System Identification: Theory and Practice*. CRC Press, 2014.

[33] Norbert Wiener. *Extrapolation, interpolation, and smoothing of stationary time series*, volume 2. MIT press Cambridge, MA, 1949.

[34] Wei Wu. Fourier transforms of stationary processes. *Proceedings of the American Mathematical Society*, 133(1):285–293, 2005.

[35] Magda Peligrad and Wei Biao Wu. Central limit theorem for fourier transforms of stationary processes. *The Annals of Probability*, pages 2009–2022, 2010.

[36] Christian P Robert and George Casella. Monte carlo integration. In *Monte Carlo Statistical Methods*, pages 71–138. Springer, 1999.

[37] Arnaud Doucet, Nando De Freitas, and Neil Gordon. An introduction to sequential monte carlo methods. In *Sequential Monte Carlo methods in practice*, pages 3–14. Springer, 2001.

[38] Valeriĭ Buldygin. *Metric characterization of random variables and random processes*.

[39] Shahar Mendelson. Discrepancy, chaining and subgaussian processes. *The Annals of Probability*, pages 985–1026, 2011.

[40] Roger L Easton. Multidimensional fourier transforms. *Fourier Methods in Imaging*, pages 325–346.

[41] Winthrop W Smith and Joanne M Smith. Handbook of real-time fast fourier transforms. *IEEE, New York*, 1995.

[42] *Yield monitor data for a corn field in Argentina with variable nitrogen*. https://rdrr.io/cran/agridat/man/lasrosas.corn.html.

[43] Rodolfo Bongiovanni and James Lowenberg-DeBoer. Nitrogen management in corn using site-specific crop response estimates from a spatial regression model. In *Proceedings of the Fifth International Conference on Precision Agriculture*, 2000.

[44] *Forest Fires Data Set, UCI Machine Learning Repository*. https://archive.ics.uci.edu/ml/datasets/Forest+Fires.

[45] Paulo Cortez and Aníbal de Jesus Raimundo Morais. A data mining approach to predict forest fires using meteorological data. 2007.