Mini-Project 1: Bag Models for Crop Yield Prediction

Harrison Zhu [1 2] **Maxime Rischard** [2] **Owen van Eer** [2] **Seth Flaxman** [1] **Dino Sejdinovic** [3]

Abstract

The central goal of crop yield predictions is to provide high-resolution mappings in order to alleviate the costs for agricultural producers to manually collect regional yield statistics and forecast future production. While recent breakthroughs in data processing of satellite imagery have provided modellers with the ability to look at the vegetation index of each individual crop field, crop production data is often only collected at lower regional resolutions, such as provincial or county-wide. Recent work on kernel methods and aggregate output Gaussian processes fit naturally into this framework. In this project, we compare different techniques that can be applied to the task of crop yield predictions and conduct experiments with a simulated Gaussian process and actual yield data in Germany, with the MODIS satellite and ECWMF weather datasets as features.

1. Introduction

The task of crop yield mapping has been recently attempted by scientists and machine learners (You et al., 2017; Lobell et al., 2015; Kuwata & Shibasaki, 2015; Bolton & Friedl, 2013; Kaneko et al.; Yang et al., 2019; Holzman & Rivas, 2016; Mladenova et al., 2017), with the aim of providing high-resolution, cheap and reliable yield mappers for the supply chain. This lends itself to the recent availability of remote sensing and geophysical data such as the Moderate Resolution Imaging Spectroradiometer (MODIS), Sentinel and European Centre for Medium-Range Weather Forecasts (ECMWF), along with corresponding advances in remote sensing processing tools and platforms such as BigQuery (Tigani & Naidu, 2014), PostGIS and Google Earth Engine (Gorelick et al., 2017). By utilising the vast amount of auxiliary data such as vegetation indices (VIs) and weather, modellers now have the data needed to model crop yield as a function of these variables. In particular, there is work that has been done to understand the correspondence between VIs and biological variables (Boryan et al., 2011).

A modern approach, as proposed by (Lobell et al., 2015) that incorporates expert knowledge would be the Scalable satellite-based Crop Yield Mapper (SCYM). However, such an approach only considers linear regression models relating synthetic yields to weather and remote-sensed variables, neglecting spatial and temporal effects between variables.

The field of spatial statistics has been widely explored by the remote sensing and geostatistics community (Cressie, 1992). The venerable technique kriging in spatial statistics is equivalent to what is known as Gaussian process (GP) regression (Rasmussen, 2003) by the machine learning community. This nonparametric technique has been successful in a variety of applied problems such as crime (Flaxman et al., 2019), elections (Flaxman et al., 2016b), remote sensing (Ton et al., 2018) and in particular epidimiology (Law et al., 2018; Bhatt et al., 2017). GPs also have attractive theoretical properties such as concentration under the lens of nonparametric Bayes' (van der Vaart et al., 2008; Vaart & Zanten, 2011).

One fundamental issue for crop yield modelling is the problem of multiple instance learning (Maron & Lozano-Pérez, 1998), where we only have access to labels at bag level but features at a higher resolution within each bag. This can be approached in 2 ways: aggregated output modelling (Tanaka et al., 2019; Law et al., 2018; Hamelijnck et al., 2019) or distribution regression (Law et al., 2017; Szabó et al., 2015; Flaxman et al., 2016a;b). While the latter has already received some attention in the context of yield mapping (Thorns, 2018; Adsuara et al., 2019; Sanchis et al., 2019; Mateo-Sanchis et al., 2019), the former has not yet been explored.

This work can be summarised by the following: First, we provide a thorough review of the GP and kernel methods in modelling bag data. We then explore aggregate GP models that can be used for crop yield modelling. Finally, we experiment 2 of our models, distribution regression and exact aggregate GP, to a synthetic and real yield dataset. To the best of our knowledge, exact aggregate GP models have not appeared in literature.

[1]Department of Mathematics, Imperial College London, London, United Kingdom [2]Cervest Limited, United Kingdom [3]Department of Statistics, Oxford University, United Kingdom. Correspondence to: Harrison Zhu <hbz15@ic.ac.uk>.

1.1. Notation

Throughout this report, we denote X as the feature space. H_k represents a reproducing kernel Hilbert space (RKHS) of functions $f : X \to \mathsf{R}$ with kernel $k : X \times X \to \mathsf{R}$, which are all defined in section 3. $\mathrm{N}_k(\mu, \ )$ denotes a k-dimensional multivariate normal distribution with mean μ and covariance matrix . For simplicity, we will drop the subscript and just use $\mathrm{N}(\mu, \ )$. X^n represents the Cartesian product of spaces $X \times X$. We assume all inner product fields to be R unless stated otherwise.

2. Gaussian Processes

2.1. Background

Gaussian processes (GPs) are the workhorse of spatial statistics. We here give a brief review of GPs and the variational GP approximation. We refer to (Rasmussen, 2003) for a detailed review.

$f f(x) g_{x \in X}$ is a Gaussian process with mean function $m : X \to \mathsf{R}$ and kernel, or covariance function, $k : X^2 \to \mathsf{R}$ if for all $X := (x_1, \ldots, x_n)^T \in X^n$,

$$f(X) := (f(x_1), \ldots, f(x_n))^T \sim N_n(m_X, k_{XX}), \quad (1)$$

where

$$m(X) := (m(x_1), \ldots, m(x_n))^T \quad (2)$$

$$k(X, X) := (k(x_i), k(x_j))_{ij}, \quad (3)$$

and we let $k(x, X) := (k(x, x_1), \ldots, k(x, x_n))$ and $k(X, x) := k(x, X)^T$ for all $x \in X$. For notation simplicity, let $K_{xX} := k(x, X), K_{Xx} := k(X, x)$, $K_{XY} := k(X, Y)$ for any $Y := (y_1, \ldots, y_m)^T \in X^m$ and $m_X := m(X)$. We denote $f \sim \mathrm{GP}(m, k)$ to represent $f f(x) g_{x \in X}$ as a GP.

Given a set of observed X with response $y \in \mathsf{R}^n$ by the GP regression $y = f(x) + \epsilon$, where $\epsilon \sim \mathrm{N}(0, \ )$ with being a diagonal matrix, the posterior distribution of f would yield for any $X \in X^m$

$$f(X \ ) | f(X), X \sim N(m_{X \ }, K_{X \ X \ }), \quad (4)$$

where m and K are the posterior mean and kernel represented by

$$m_{X \ } := m_X \ + K_{X \ X}[K_{XX} + \ ](y \ \ m_X), \quad (5)$$

$$K_{X \ X \ } := K_{X \ X} \ \ K_{X \ X}[K_{XX} + \ ]^{-1} K_{XX} \ . \quad (6)$$

However, inverting the $n \times n$ matrix $K_{XX} +$ requires $O(n^3)$ operations and $O(n^2)$ storage, which makes it computationally infeasible for large datasets. Methods for approximating GP regression make it possible to obtain $O(nm^2)$ via inducing points in the function evaluation space, using methods such as the Nyström method (Williams & Seeger, 2001) or Random Fourier Features (Rahimi & Recht, 2008).

2.2. Variational Learning of GPs

A variational approach is to learn a variational distribution of the posterior distribution of the inducing points by maximising the log-marginal likelihood by maximising its lower bound (Titsias, 2009).

Suppose we have a set of inducing points $(u_l)_{l=1}^m$ that are given by $u_l := f(w_l)$ at landmark points $W := (w_1, \ldots, w_m)^T \in X^n$. The inducing points can be picked at random, or via k-means clustering. The distribution $p(u|W)$ is thus

$$u|W \sim \mathrm{N}(m_W, K_{WW}). \quad (7)$$

With kernel and mean function parameters θ, we thus have

$$\begin{aligned} \log p(y|\theta) &= \log \int \int p(y, f, u|X, W, \theta) df \, du \\ &\int \int \log \left[p(y|f, \theta) \frac{p(u|W)}{q(u)} \right] p(f|u, \theta) q(u) df \, du \\ &= \int \int \log \left[p(y|f, \theta) \right] p(f|u, \theta) q(u) df \, du \\ &\mathrm{KL}(q(u) \| p(u|W)) := L(q, \theta), \end{aligned} \quad (8)$$

where we used Jensen's inequality, $q(u) = \mathrm{N}(\eta_u, \ _u)$ is the variational distribution that we optimise by maximising the evidence lower bound (ELBO) $L(q, \theta) = L(\eta_u, \ _u, \theta)$ with gradient-based methods and KL($\|$) is the Kullback-Leibler divergence.

$q(u)$ is an approximation to the posterior $p(u|y)$, which then yields the posterior approximation of f as $q(f) = \int p(f|u) q(u) du$. Both terms in the integrand are Gaussian and so the resulting integral is also Gaussian.

2.3. Aggregate Output GPs

GPs can be used to solve multiple instance learning problems by defining the aggregate output GP (Law et al., 2018).

We first define the *bag data* to be the set $f(f x_i^a g_{i=1}^{N_a}, y^a) g_{a \in I}$, where x_i^a and y^a are the bag features and labels, with bag indices I so that $|I| = n$. A *bag* is $(f x_i^a g_{i=1}^{N_a}, y^a)$, which is a collection of features with bag index a. Denote the bag features as $X^a = (x_1^a, \ldots, x_{N_a}^a)^T$ and all the features as $X = (X_1, \ldots, X_n)^T$.

Suppose that

$$y_i^a | x_i^a \sim \mathrm{N}\left(w_i^a \mu(x_i^a), (w_i^a)^2 (\tau_i^a)^2 \right), \quad (9)$$

where $\mu(\ )$ is an unspecified mean function. Note that we have $|A^a| = \sum_{i=1}^{N_a} |A_i^a|$. We give $\mu(\ )$ a GP prior f

GP(m, k) and obtain the aggregated distribution

$$y^a = \sum_{i=1}^{N_a} y_i^a \, jf x_i^a g_{i=1}^{N_a} \tag{10}$$

$$N\left(\sum_{i=1}^{N_a} w_i^a \mu(x_i^a), \sum_{i=1}^{N_a} (w_i^a \tau_i^a)^2\right). \tag{11}$$

Suppose our training set is summarised in (X, y), where $y = (y_i^a)_{a;i}$ are the latent observations. Similarly, we have the same setting for the test set $(X \ , y \ )$. In vector form, we have

$$\begin{pmatrix} y \\ f \end{pmatrix} = \begin{pmatrix} W & 0 \\ 0 & W \end{pmatrix} \begin{pmatrix} y \\ f \end{pmatrix},$$

where $W \ 2 \ \mathsf{R}^n \ \sum_{a=1}^n N_a$ and $W \ 2 \ \mathsf{R}^n \ \sum_{a=1}^n N_a$ are weight matrices for the training and prediction set. The training weights correspond to

$$W := \begin{pmatrix} (w_1^1, \ldots, w_{N_1}^1) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & (w_1^n, \ldots, w_{N_n}^n) \end{pmatrix},$$

and similarly for the test weights $W \ $.

For the latter, we have the distribution

$$\begin{pmatrix} y \\ f \end{pmatrix} \ \ \mathrm{N}\left(0, \begin{pmatrix} K_{XX} + & K_{XX} \\ K_{X \ X} & K_{X \ X} \end{pmatrix}\right),$$

where $\ := \mathrm{diag}(\tau_i^a)$, giving

$$\begin{pmatrix} y \\ f \end{pmatrix} \ \ \mathrm{N}\Bigg(0, \begin{pmatrix} W(K_{XX} + \ )W^T & WK_{XX} \ W^{\ T} \\ W \ K_{X \ X}W^T & W \ K_{X \ X} \ W^{\ T} \end{pmatrix}\Bigg)$$

The posterior thus follows as

$$f \ jy, f, X \ \ \mathrm{N}(m \ (X \ ), V \ (X \ )), \tag{12}$$

where

$$m \ (X \ ) = W \ K_{X \ X}W^T[W(K_{XX} + \ )W^T] \ ^1 y, \tag{13}$$

$$V \ (X \ ) = W \ K_{X \ X} \ W^{\ T} \ \ W \ K_{X \ X}W^T \ [W(K_{XX} + \ )W^T] \ ^1 WK_{XX} \ W^{\ T}. \tag{14}$$

Storing the weight and feature matrices is expensive, and so in practice we compute the resulting weighted matrices by batch.

The marginal likelihood, with hyperparameters θ, can thus be written as

$$L(\theta) \ \ \frac{1}{2} y^T [W(K_{XX} + \ )W^T] \ ^1 y \ \ \frac{1}{2} \log jW(K_{XX} + \ )W^T j. \tag{15}$$

In addition, we can optimise the marginal likelihood by optimising the marginal likelihood via gradient descent.

2.4. Spatially-aggregated GPs

GPs that consider distributions over areas or regions rather than points in space and time have been explored by the geostatistics community in the context of muliresolution dynamic space time models (DSTMs) (Cressie, 1992). For the sake of crop yield mapping, the datasets are often associated to provinces and counties, making it a necessity to consider distributions over areas.

We first define an areal data, similar to a bag data, as $f(A^a, y^a)g_{a2I}$, where A^a is a region.

Recently, there has been work on spatially-aggregated GPs (Tanaka et al., 2019) where we define L latent independent GPs

$$g_l \ \ \mathrm{GP}(\nu_l, \gamma_l), \tag{16}$$

where $\nu_l : X \ ! \ \mathsf{R}$ and $\gamma_l : X^2 \ ! \ \mathsf{R}$. Now define the regression

$$f(x) = w^T g(x) + \epsilon(x), \tag{17}$$

where $w \ 2 \ \mathsf{R}^L$ is a weight vector, $g(x) := (g_1(x), \ldots, g_L(x))^T$, $\epsilon \ \ \mathrm{GP}(0, \lambda)$ is a GP noise and $\lambda : X^2 \ ! \ \mathsf{R}$. This is equivalent to

$$f(x) \ \ \mathrm{GP}(m, k), \tag{18}$$

with

$$m(x) := w^T \nu(x) \tag{19}$$

$$k(x, x^0) = w^T \ (x, x^0)w + \lambda(x, x^0), \tag{20}$$

where $\nu(x) := (\nu_1(x), \ldots, \nu_L(x))^T$ and $(x, x^0) := \mathrm{diag}(\gamma_1(x, x^0), \ldots, \gamma_L(x.x^0))^T$.

To account for spatial aggregation, we model the areal labels $y = (y^1 \ldots, y^n)^T \ 2 \ \mathsf{R}^n$ with the likelihood

$$yjf(x) \ \ \mathrm{N}(\mu(x), \ ), \tag{21}$$

where $A(\ )$ is the measure of spatial aggregations on X, $:= \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ and

$$\mu(x) = \begin{pmatrix} \int_{A_1} f(x)dA(x)/A(A_1) \\ \vdots \\ \int_{A_n} f(x)dA(x)/A(A_n) \end{pmatrix}. \tag{22}$$

The kernel is of the form

$$K = \left(\int_{A_i}\int_{A_j} k(x, x^0)\frac{dA(x)dA(x^0)}{A(A_1)A(A_n)}\right)_{i;j \, 2I} + \ . \tag{23}$$

Given a new point $x \ \ 2 \ X$, we can also write down the kernel evaluation as

$$k(x \ , \bigcup_{a=1}^n A_a) := \left(\int_{A_1} k(x \ , x)\frac{dA(x)}{A(A_1)}, \ldots, \int_{A_n} k(x \ , x)\frac{dA(x)}{A(A_n)}\right). \tag{24}$$

Evaluation of equations 22 and 24 are usually intractable, and so a Monte Carlo integration or quadrature approach can be applied. In addition, $A(\ )$ is usually taken to be fixed for each A_a, for example population (Law et al., 2018) or areal proportions (Tanaka et al., 2019).

The rest of the inference steps follows from section 2.1

2.5. Variational Learning on Aggregated Output GPs (VBagg)

The variational way to do inference on the aggregate output model in section 2.3 would be to optimise the ELBO, which can be written as

$$\begin{aligned} &L(\eta_u, \quad _u, W, \theta) = \\ &\quad \frac{1}{2}\sum_{a=1}^{n}\left\{\frac{(y^a)^2 \quad 2y^a\mathbf{1}^T\boldsymbol{m}^a + \mathbf{1}^T(\boldsymbol{S}^a + \boldsymbol{m}^a(\boldsymbol{m}^a)^T)\mathbf{1}}{N_a\tau_a}\right\} \\ &\quad \frac{1}{2}\sum_{a=1}^{n}\log(2\pi N_a\tau_a) \quad \mathrm{KL}(q(u)jjp(ujW)), \end{aligned} \tag{25}$$

where

$$\boldsymbol{m}^a := m_{X^a} + K_{X^aW}K_{WW}^{1}(\eta_u \quad \mu_W), \tag{26}$$

$$\begin{aligned} \boldsymbol{S}^a := &K_{X^aX^a} \quad K_{X^aW}(K_{WW}^{1} \\ &K_{WW}^{1} \quad _uK_{WW}^{1})K_{WX^a}. \end{aligned} \tag{27}$$

The derivation can be found in (Law et al., 2018).

3. Kernel Methods

3.1. Background

We hereby give a brief review of kernel methods, a function-space-based set of methods that involves a special Hilbert space that is induced by a kernel, called the reproducing kernel Hilbert space (RKHS), to learn feature representations of the feature space. We refer to (Muandet et al., 2017) for a thorough review.

Suppose $\varphi : X \ ! \ H$ is a map to a high-dimensional feature space (a Hilbert space) H. Then we can define a kernel as

$$k(x, x^0) := h\varphi(x), \varphi(x^0)i_H. \tag{28}$$

All kernels are positive definite, in the sense that the kernel matrix K_{XX} formed by $X = (x_1 \ldots, x_n)^T \ 2 \ X^n$ is positive definite, since

$$\begin{aligned} \alpha^T K_{XX}\alpha &= \sum_{i;j=1}^{n}\alpha_i\alpha_j k(x_i, x_j) \\ &= \sum_{i;j=1}^{n}\alpha_i\alpha_j h\varphi(x_i), \varphi(x_j)i_H \\ &= \left\|\sum_{i=1}^{n}\alpha_i\varphi(x_i)\right\|_H \quad 0, \end{aligned}$$

for all $\alpha = (\alpha_1, \ldots, \alpha_n)^T \ 2 \ \mathsf{R}^n$.

Well-known properties of kernels include being closed under addition and multiplication. Furthermore, kernels can also be defined on general Hilbert spaces H so that for $f, g \ 2 \ H$, we have the linear kernel

$$k(f, g) = hf, gi_H. \tag{29}$$

A reproducing kernel Hilbert space (RKHS) H_k is a Hilbert space of functions $f : X \ ! \ \mathsf{R}$ equipped with the reproducing kernel $k : X^2 \ ! \ \mathsf{R}$ so that

- $k(\ , x) \ 2 \ H_k$ for all $x \ 2 \ X$,
- $hf, k(\ , x)i_H = f(x)$ for all $x \ 2 \ X, f \ 2 \ H_k$. This property is known as the reproducing property.

In particular, the Moore-Aronzajn theorem guarantees that for each positive definite kernel k there exists a unique RKHS with k being the corresponding kernel.

3.2. Distribution Regression

Define the kernel mean embedding of a probability measure P on X as

$$\mu_{\mathbb{P}} := \int_X k(\ , x)\mathsf{P}(dx) \ 2 \ H_k, \tag{30}$$

where it is embedded onto the RKHS H_k with kernel k. Similarly, the emperical mean embedding is

$$\hat{\mu}_{\mathbb{P}} := \mu_{\hat{\mathbb{P}}} = \int_X k(\ , x)\hat{\mathsf{P}}(dx) = \frac{1}{n}\sum_{i=1}^{n}k(\ , x_j), \tag{31}$$

for a set of points $fx_jg_{j=1}^{n} \ ^{iid} \ \mathsf{P}$. Using the reproducing property of the RKHS, we obtain the identity for 2 probability measures $\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{P}_j}$

$$h\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{P}_j}i_{H_k} = \frac{1}{N_iN_j}\sum_{l=1}^{N_i}\sum_{r=1}^{N_j}k(x_l^i, x_r^j). \tag{32}$$

Suppose we have a bag data $f(fx_i^ag_{i=1}^{N_a}, y^a)g_{a2I}$ and the feature map $\varphi(x) = k(\ , x) \ 2 \ H_k$, let the empirical mean embedding be

$$\hat{\mu}^a := \sum_{i=1}^{N_a}\varphi(x_i^a). \tag{33}$$

Distribution regression (Szabó et al., 2015) culminates to minimising the loss

$$L(f) := \frac{1}{n}\sum_{a=1}^{n}(y_i \quad f(\hat{\mu}^a))^2 + \lambda kfk_{H_K}^2, \tag{34}$$

with $K : H_k \quad H_k \; ! \; \mathbb{R}$ being a "second-level" kernel on the RKHS and over $f \; 2 \; H_K$ and λ being the regularisation parameter. With a linear kernel we would get the ridge regression form on mean embeddings (Szabó et al., 2015), although other kernels such as the RBF kernel has also been used to improve performance (Muandet et al., 2017).

By the Representer theorem and reproducing property (Szabó et al., 2015), this has a unique solution $f = \sum_{j=1}^{n} \alpha_j K(\,, \hat{\mu}^j)$ with

$$\alpha = (\alpha_1, \ldots, \alpha_n)^T = (\boldsymbol{K} + \lambda I_n)^{\;1} \boldsymbol{y}, \tag{35}$$

where $\boldsymbol{K}_{ij} = K(\hat{\mu}^i, \hat{\mu}^j)$ and $\boldsymbol{y} = (y^1 \ldots, y^n)^T$. Finally, a prediction on a new bag with features $\{x_i\}_{i=1}^{N}$ would be computed with

$$\begin{aligned} \hat{y} &= f(\hat{\mu}) = \sum_{a=1}^{n} \alpha_a K(\hat{\mu}, \hat{\mu}^a) \\ &= K \wedge \boldsymbol{M} \alpha, \end{aligned} \tag{36}$$

where $\boldsymbol{M} = (\hat{\mu}^1, \ldots, \hat{\mu}^n)^T$.

The kernel parameters and regularisation parameter λ can be tuned via cross validation, but this is expensive given that distribution regression requires $O(N_i N_j)$ operations for the evaluation of each entry in $\boldsymbol{K}$, which is of size $O(n^2)$.

3.3. Bayesian Distribution Regression

Bayesian distribution regression (Law et al., 2017) is way to modify distribution regression with uncertainty quantification due to bag sizes via kernel mean priors (Law et al., 2017) and Bayesian linear regression.

We first suppose we have a set of landmark points $(u_l)_{l=1}^{m}$ as of section 2.2. We modify the original feature map φ to

$$\varphi(x) = (k(x, u_1), \ldots, k(x, u_m))^T, \tag{37}$$

and have the emperical mean embedding $\hat{\mu}^a(u) = \frac{1}{N_a} \sum_{i=1}^{N_a} \varphi(x_i^a)$.

Then we define a prior over the mean embeddings (Flaxman et al., 2016a) as

$$\mu \quad \mathrm{GP}(m_0, \eta r(\,, \,)), \tag{38}$$

where $\eta \; 2 \; \mathbb{R}$, r is chosen so that $\mu \; 2 \; H_k$ almost surely and that for each bag $a \; 2 \; I$

$$\hat{\mu}^a(u) | \mu^a(u) \quad \mathrm{N}(\mu^a(u), \frac{\;^a}{N_a}), \tag{39}$$

where $\;^a$ is the covariance matrix for the bag a. The expression suggests that there is more uncertainty for bags with fewer items. Using the theory for 2.1, we have a GP over $\mu^a | \hat{\mu}^a$, which yields a posterior distribution for $h \; 2 \; X^d$

$$\mu^a(h) | \hat{\mu}^a(u) \quad \mathrm{N}(r^a, Q^a), \tag{40}$$

where

$$r^a = R_{hu}(R_{uu} + \frac{\;^a}{N^a})^{\;1} (\hat{\mu}^a(u) \quad m_0, \tag{41}$$

$$Q^a = R_{hh} \quad R_h(R_{uu} + \frac{\;^a}{N^a})^{\;1} R_{hu^T}, \tag{42}$$

with $(R_{uu})_{ij} = \eta r(u_i, u_j)$, $(R_{uh})_{ij} = \eta r(u_i, h_j)$ and $(R_h)_{ij} = \eta r(h_i, h_j)$.

Suppose we now model the bag response using

$$y^a | f, \mu^a \quad \mathrm{N}(h f \,, \mu^a i_{H_k}, (\tau^a)^2). \tag{43}$$

Taking a set of points $h \; 2 \; X^d$ and using the Representer approximation $f \quad \sum_{i=1}^{d} \alpha_i k(\,, h_i)$, we obtain the approximate likelihood

$$y^a | \mu^a(h), \alpha \quad \mathrm{N}(\alpha^T \mu^a(h), (\tau^a)^2). \tag{44}$$

Integrating out μ^a and using the properties of the integration of 2 Gaussians, we get the bag label response conditioned on the empirical mean embeddings as

$$y^a | \hat{\mu}^a(u), \alpha \quad \mathrm{N}(\alpha^T r^a, \alpha^T Q^a \alpha + (\tau^a)^2). \tag{45}$$

A prior distribution can be placed on α, having the effect of regularisation of f, similar to that in the canonical distribution regression with $\|f\|_{H_K}$. In addition, the points u and h can be picked by k-means clustering, similar to the variational GP in section 2.2.

It is then possible to either use MAP estimation or Markov chain Monte Carlo (MCMC) algorithms such as Hamiltonian Monte Carlo (HMC) (Law et al., 2017) to infer the model parameters.

3.4. Link between Bayesian distribution regression and aggregate output models

There is in fact a very close relationship between Bayesian distribution regression and aggregate output models that we defined in the previous sections. Notably, both methods deal with the Gaussian observation model of the bag labels. Both models specify a similar likelihood, one using a linear function of the mean embeddings and the other uses an aggregate of the underlying GP evaluations at bag item level. The prior in the aggregate GP method is a GP and it is the mean embeddings for Bayesian distribution regression, meaning that these are actually just 2 similar ways of aggregating to the bag level. There has been work already that has justified the connection between GP regression and distribution regression with the linear kernel (Kanagawa et al., 2018), with the solution to both methods being mathematically identical. Therefore we would expect similar results from these 2 families of methods.

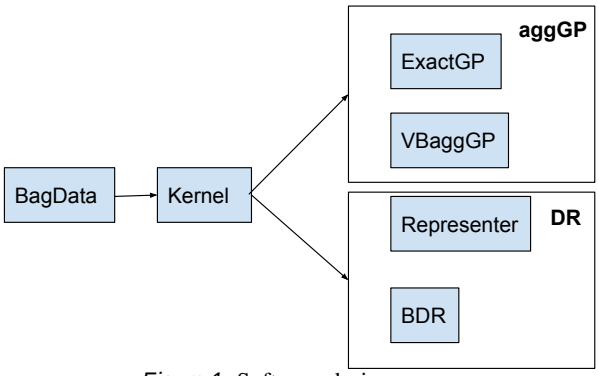

Figure 1. Software design.

4. Software

We implemented several classes that can be used for the next steps of this project. Notably, we have built from scratch classes `Kernel`, `BagData`, `Representer`, `BDR`, `ExactGP` and `VBaggGP`. Due to time constraints, we have not been able to finish `BDR` and `VBaggGP` but hope to complete these as the next step for the project. We have also not implemented the spatially-aggregated GP in section 2.4, which we leave for future work.

The aim is to develop these further into reproducible and unit-tested Python libraries for collaborative research. The design model is summarised in Figure 1. The code is available upon request.

5. Experiments

We experiment on 2 datasets using both the exact aggregate output GP (which we call ExactGP) and distribution regression (DR). Due to time constraints, we were not able to compute the marginal likelihood of the GP model. For the distribution regression model, we use the linear kernel, giving us the kernel ridge regression method. We conducted our experiments on the Cervest high performance computing CPU cluster with 1TB RAM.

5.1. Synthetic GP Example

We generate a bag data that exists on a 2D grid using the data generating process

$$y^a = \sum_{i=1}^{N^2} w_i^a y_i^a, \tag{46}$$

$$y_i^a \quad \mathrm{N}(f(x_i^a), \tau_i^a), \quad \forall i = 1, \ldots, N^2, \tag{47}$$

$$f \quad \mathrm{GP}(0, k), \tag{48}$$

where the points x_i^a are picked uniformly by discretising $X = [\ 1, 1]^2$ into a uniform grid with using length and width step sizes of N. The kernel is picked to be the standard

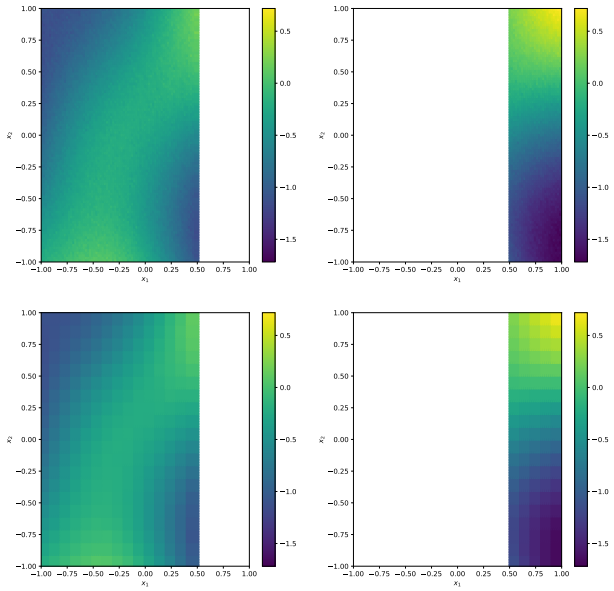

Figure 2. Gaussian synthetic grid generated. Top being the latent response and the bottom being the bag response, left being the training set and right being the test set.

RBF kernel

$$k(x, y) = \exp\left(\frac{jjx \quad yjj^2}{2\ell^2} \right), \tag{49}$$

with lengthscale parameter ℓ. For our experiments, we set $w_i^a = N^{\ 1}$, $N = 5$, $\tau_i^a = 0.01$, $\lambda = 0.01$ and $\ell = 1$, giving us 400 bags, each with 25 bag items. Due to time constraints, we did not tune our parameters.

Figure 2 shows the train test split when the training and test sets are set to 300 and 100 bags respectively. We can see that the predictions are quite good on this simple dataset, as shown by Figure 3 and 4 and Table 1. Furthermore, we were also able to propagate bag uncertainty for ExactGP, making it useful to identify where the model is most uncertain about its predictions, as shown in Figure 5.

Table 1. Prediction metrics for the synthetic dataset.

Model	RMSE	MAPE
DR	0.1201	0.0810
ExactGP	0.1201	0.0809

5.2. Yield Prediction for Germany with remote sensing

Our data contains wheat yield data from 2002 to 2017 in Germany over 13 NUTS 1 (a European statistical region classification) regions, as shown in Figure 6. Each bag will

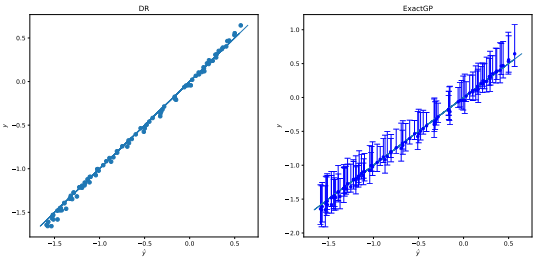

Figure 3. Prediction results with actual labels y and predictions $\hat{y}$ with distribution regression and ExactGP plotted with 1.96*σ credible intervals.

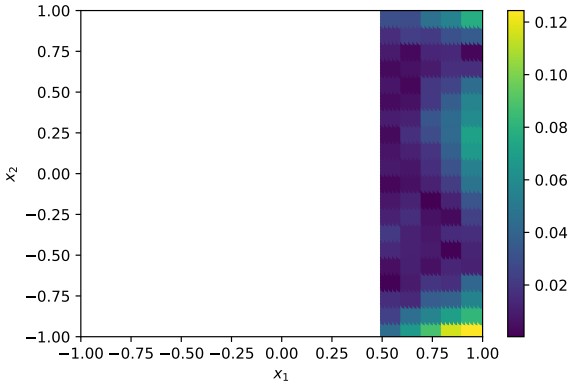

Figure 4. Prediction results with absolute error $|y - \hat{y}|$ with distribution regression. The results for ExactGP were similar.

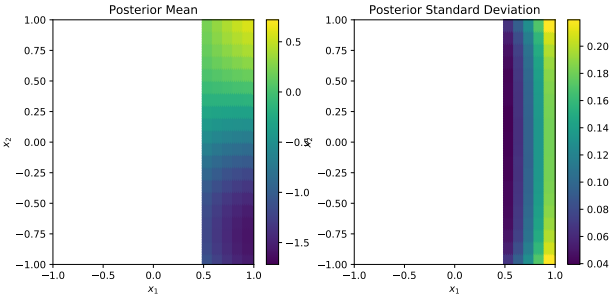

Figure 5. Prediction results with ExactGP with posterior standard deviation.

be a NUTS 1 region for a particular year with the corresponding NUTS 3 regions at different days being the bag items. Figure 7 show the number of NUTS 3 regions in each NUTS 1 region. The corresponding yield and production are shown in Figures 8 and 9.

Our features are remote sensing data from the MODIS13Q1 dataset, which an open source dataset of MODIS (Moderate Resolution Imaging Spectroradiometer) released by NASA, a component of the Earth Observing System (EOS) programme. We will use the vegetation index called Normalised Difference Vegetation Index (NDVI) (Huete et al., 2002; Rouse Jr et al., 1974), as defined by

$$\text{NDVI} = \frac{\rho_{\text{NIR}} \quad \rho_{\text{Red}}}{\rho_{\text{NIR}} + \rho_{\text{Red}}}, \tag{50}$$

where ρ_{NIR} and ρ_{Red} the percentage reflectance (Viña et al.,

Figure 6. A map illustrating the different NUTS 1 regions that we use as bags, as colour-coded. Picture taken from https://www.czso.cz/csu/czso/2-1373-05--08.

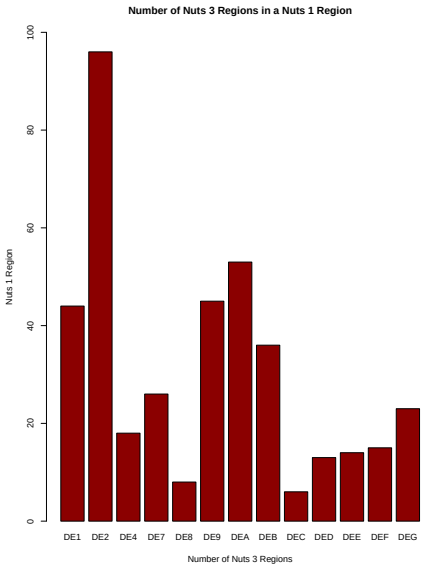

Figure 7. Number of NUTS 3 regions in each NUTS 1 region in Germany.

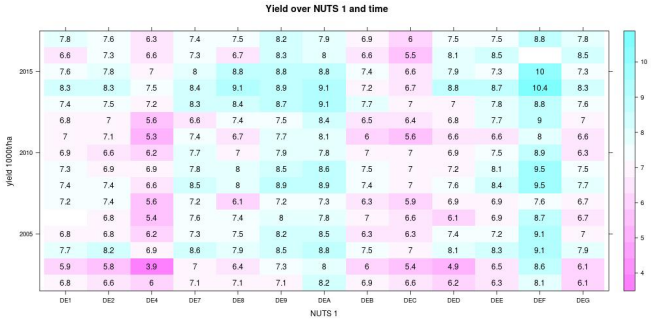

Figure 8. A table of the yield recorded for each NUTS 1 region over 2003-2017.

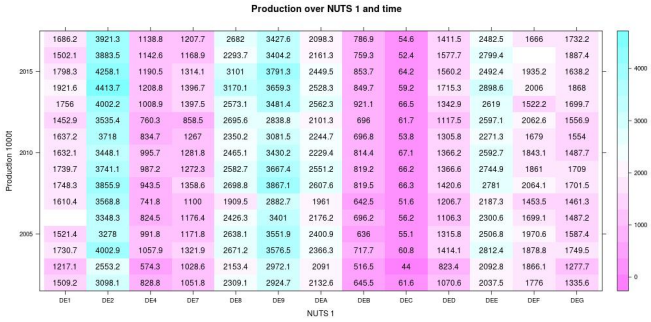

Figure 9. A table of the production recorded for each NUTS 1 region over 2003-2017.

2011) calculated at red and NIR spectral bands. For each point on earth, MODIS makes a record every 16 days at 250m 250m resolution (for the highest resolution). For example, Figure 11 shows a time series plot of the NDVI over a region.

We also have another feature data that comes from The European Centre for Medium-Range Weather Forecasts (ECMWF), which contains features: leaf area index, volumetric soil water layers 1-4, soil temperature level 1-4 and 2 metre temperature. However, this is only available for 2014-2017.

With regards to the yield data, it is access from the Eurostat[1] database. The feature datasets are prepared in collaboration with Cervest Ltd and are not included along with the paper.

Each bag contains N_a bag items that come from NUTS 3 regions for different days of the year, and for the features each bag item uses the mean of all the pixel features within the bag item. Our features are the triple (x, y, NDVI), with (x, y) being the average latitude and longitude of the corresponding NUTS 3 region. We do a train-test split with years 2006-2014 being our training set and years 2015-2017 being our test set, giving us 116 training bags an 38 test bags, with a total of 78947 and 26742 items respectively.

We set $w_i^a = N_a$ [1]. We also set the kernel to the RBF kernel again with bandwidth $\ell = 1$. Again, we did not do any cross-validation or gradient descent to tune the parameters due to time constraints. The bag response noise was neglected, although in future experiments this should be included.

We can see that both methods perform identically from Table 2 and Figure 10. The RMSE results are competitive to results from other models produced internally at Cervest Ltd, and the next step would be to validate this result on the same dataset.

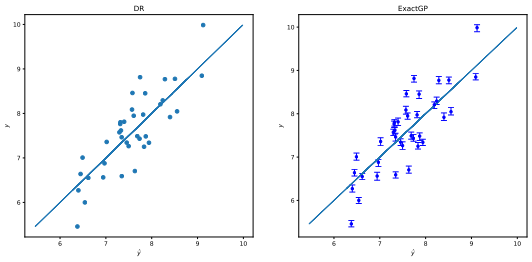

Figure 10. Prediction results with actual labels y and predictions $\hat{y}$ with distribution regression and ExactGP plotted with 1.96*σ credible intervals.

Table 2. Prediction metrics for the yield dataset.

Model	RMSE	MAPE
DR	0.4940	0.0555
ExactGP	0.4940	0.0555

6. Conclusion and Future Work

We reviewed GP models for aggregate output response and spatially aggregated datasets, introducing the exact aggregate output GP, spatially aggregated GP and variational aggregate output GP. Then, we motivated distribution regression and explored Bayesian distribution regression. We implemented a suite of software that will be further developed in the future and experiment on synthetic and actual crop yield data from Germany. Our results showed that distribution regression and exact aggregate output GPs perform well on both datasets, and in fact yield very similar results. Another important final step to evaluate performance would be to look at cross-validation metrics, such as spatial cross-validation (Brenning, 2012).

Our next step would be to fix the marginal likelihood computation in the software for the GP classes, complete the other model classes and unit test all the components on the synthetic dataset.

There are several ways to improve our current modelling

[1] https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=apro_cpsh1&lang=en

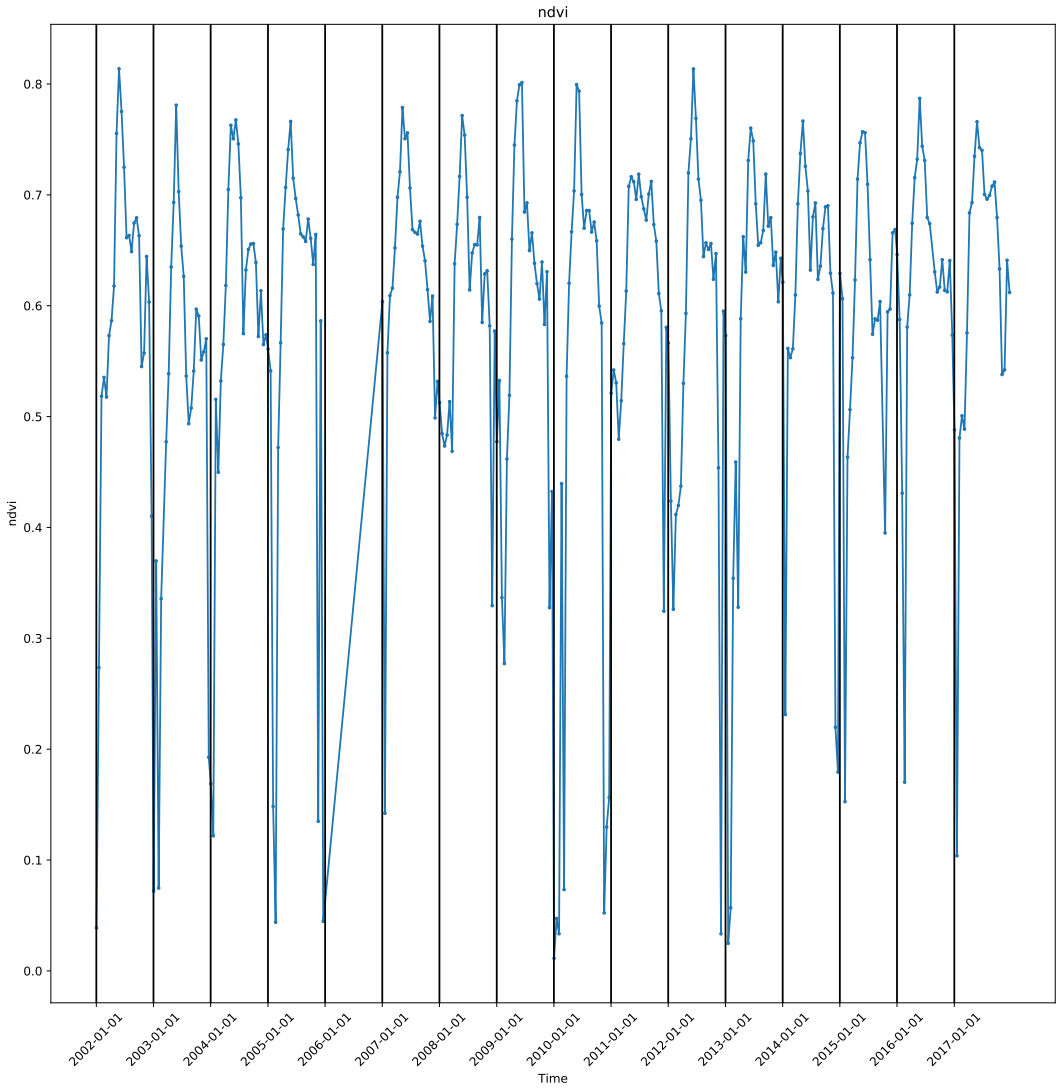

Figure 11. A time series plot of the NDVI over DE11A NUTS 3 region from 2002-2017.

strategy. Firstly, we can build a model on data from the US, which is more abundant, and use transfer learning as detailed in (Tanaka et al., 2019) to Europe. As suggested in (Thorns, 2018) and (Adsuara et al., 2019), there is a way to do distribution regression and BDR on multiple feature datasets - known as the multisource method. This could potentially be possible for the GP, and needs to be explored. Another issue that we ran into were the computational and storage complexity of distribution regression and aggregate output GP, as detailed in sections 2 and 3. This could be resolved by existing methods such as RBF networks and BDR (Law et al., 2017) or variational approaches to aggregate output GPs (Law et al., 2018) and recent advances in efficient exact GP inference (Wang et al., 2019). In particular, these Bayesian approaches would allow us to be able to estimate the distribution of the response at item or individual level, making it desirable for applications such as yield optimisation or yield surveying.

In terms of feature and representation learning of our feature datasets, potential approaches could be to first estimate or obtain the start and end of season growths, such as via the SMAP dataset (Thorns, 2018). These approaches can be

tackled by perhaps first a segmentation of all the farms via methods such as the U-Net (Ronneberger et al., 2015), extract time series of the individual field level features and then building a crop classification model. Once this is done, a spatiotemporal change-point detection algorithm can be applied to detect the start and end of seasons.

The problem of crop yield prediction is that the number of responses we have is very low, as they are usually only report on a annual basis. Recent advances in few-shot learning models such as the Neural Process family (Kim et al., 2019; Garnelo et al., 2018b;a; Gordon et al., 2019). There has been unpublished experiments of the Neural Process family on synthetic 1D and 2D Gaussian processes and they are shown to perform very well and competitive to exact GPs. Crucially, the computational complexity of the Neural Process family is significantly lower than exact GPs and since the weights are trained via gradient descent, it does not suffer from the numerical instabilities of inverting the GP kernel matrix.

Lastly, one of the ideas that we would like to motivate is the link between aggregation and attention, which has gained interests in natural language processing (Vaswani et al., 2017), image processing (Parmar et al., 2018) and neural processes (Kim et al., 2019). We only looked at linear fixed aggregation weights for our models, but these could potentially be generalised via attention layers.

References

Adsuara, J. E., Pérez-Suay, A., Muñoz-Marí, J., Mateo-Sanchis, A., Piles, M., and Camps-Valls, G. Nonlinear distribution regression for remote sensing applications. *IEEE Transactions on Geoscience and Remote Sensing*, 57(12):10025–10035, 2019.

Bhatt, S., Cameron, E., Flaxman, S. R., Weiss, D. J., Smith, D. L., and Gething, P. W. Improved prediction accuracy for disease risk mapping using gaussian process stacked generalization. *Journal of The Royal Society Interface*, 14(134):20170520, 2017.

Bolton, D. K. and Friedl, M. A. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. *Agricultural and Forest Meteorology*, 173:74–84, 2013.

Boryan, C., Yang, Z., Mueller, R., and Craig, M. Monitoring us agriculture: the us department of agriculture, national agricultural statistics service, cropland data layer program. *Geocarto International*, 26(5):341–358, 2011.

Brenning, A. Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The r package sperrorest. In *2012 IEEE international geoscience and remote sensing symposium*, pp. 5372–5375. IEEE, 2012.

Cressie, N. Statistics for spatial data. *Terra Nova*, 4(5): 613–617, 1992.

Flaxman, S., Sejdinovic, D., Cunningham, J. P., and Filippi, S. Bayesian learning of kernel embeddings. In *UAI*, 2016a.

Flaxman, S., Sutherland, D. J., Wang, Y.-X., and Teh, Y. W. Understanding the 2016 us presidential election using ecological inference and distribution regression with census microdata. 2016b.

Flaxman, S., Chirico, M., Pereira, P., and Loeffler, C. Scalable high-resolution forecasting of sparse spatiotemporal events with kernel methods: A winning solution to the nij "real-time crime forecasting challenge". *Ann. Appl. Stat.*, 13(4):2564–2585, 12 2019. doi: 10.1214/19-AOAS1284. URL https://doi.org/10.1214/19-AOAS1284.

Garnelo, M., Rosenbaum, D., Maddison, C. J., Ramalho, T., Saxton, D., Shanahan, M., Teh, Y. W., Rezende, D. J., and Eslami, S. Conditional neural processes. *arXiv preprint arXiv:1807.01613*, 2018a.

Garnelo, M., Schwarz, J., Rosenbaum, D., Viola, F., Rezende, D. J., Eslami, S., and Teh, Y. W. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018b.

Gordon, J., Bruinsma, W. P., Foong, A. Y. K., Requeima, J., Dubois, Y., and Turner, R. E. Convolutional conditional neural processes, 2019.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 2017. doi: 10.1016/j.rse.2017.06.031. URL https://doi.org/10.1016/j.rse.2017.06.031.

Hamelijnck, O., Damoulas, T., Wang, K., and Girolami, M. Multi-resolution multi-task gaussian processes, 2019.

Holzman, M. E. and Rivas, R. E. Early maize yield forecasting from remotely sensed temperature/vegetation index measurements. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(1): 507–519, 2016.

Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X., and Ferreira, L. G. Overview of the radiometric and biophysical performance of the modis vegetation indices. *Remote sensing of environment*, 83(1-2):195–213, 2002.

Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

Kaneko, A., Kennedy, T., Mei, L., Sintek, C., Burke, M., Ermon, S., and Lobell, D. Deep learning for crop yield prediction in africa.

Kim, H., Mnih, A., Schwarz, J., Garnelo, M., Eslami, A., Rosenbaum, D., Vinyals, O., and Teh, Y. W. Attentive neural processes. *arXiv preprint arXiv:1901.05761*, 2019.

Kuwata, K. and Shibasaki, R. Estimating crop yields with deep learning and remotely sensed data. In *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 858–861. IEEE, 2015.

Law, H. C., Sejdinovic, D., Cameron, E., Lucas, T., Flaxman, S., Battle, K., and Fukumizu, K. Variational learning on aggregate outputs with gaussian processes. In *Advances in Neural Information Processing Systems*, pp. 6081–6091, 2018.

Law, H. C. L., Sutherland, D. J., Sejdinovic, D., and Flaxman, S. Bayesian approaches to distribution regression. In *AISTATS*, 2017.

Lobell, D. B., Thau, D., Seifert, C., Engle, E., and Little, B. A scalable satellite-based crop yield mapper. *Remote Sensing of Environment*, 164:324–333, 2015.

Maron, O. and Lozano-Pérez, T. A framework for multiple-instance learning. In *Advances in neural information processing systems*, pp. 570–576, 1998.

Mateo-Sanchis, A., Piles, M., Muñoz-Marí, J., Adsuara, J. E., Pérez-Suay, A., and Camps-Valls, G. Synergistic integration of optical and microwave satellite data for crop yield estimation. *Remote Sensing of Environment*, 234:111460, 2019.

Mladenova, I. E., Bolten, J. D., Crow, W. T., Anderson, M. C., Hain, C. R., Johnson, D. M., and Mueller, R. Intercomparison of soil moisture, evaporative stress, and vegetation indices for estimating corn and soybean yields over the us. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(4):1328–1343, 2017.

Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends R in Machine Learning*, 10(1-2):1–141, 2017.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, Ł., Shazeer, N., Ku, A., and Tran, D. Image transformer. *arXiv preprint arXiv:1802.05751*, 2018.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pp. 1177–1184, 2008.

Rasmussen, C. E. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pp. 63–71. Springer, 2003.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Rouse Jr, J., Haas, R., Schell, J., and Deering, D. Monitoring vegetation systems in the great plains with erts. 1974.

Sanchis, A. M., Adsuara, J., Piles, M., Perez-Suay, A., Muñoz-Marí, J., and Camps-Valls, G. Multisensor distribution regression for crop yield estimation. In *Geophysical Research Abstracts*, volume 21, 2019.

Szabó, Z., Gretton, A., Póczos, B., and Sriperumbudur, B. Two-stage sampled learning theory on distributions. In *Artificial Intelligence and Statistics*, pp. 948–957, 2015.

Tanaka, Y., Tanaka, T., Iwata, T., Kurashima, T., Okawa, M., Akagi, Y., and Toda, H. Spatially aggregated gaussian processes with multivariate areal outputs. In *Advances in Neural Information Processing Systems*, pp. 3000–3010, 2019.

Thorns, D. Distribution Regression for Crop Yield Prediction. Master's thesis, Department of Statistics, Oxford University, United Kingdom, 2018.

Tigani, J. and Naidu, S. *Google BigQuery Analytics*. John Wiley & Sons, 2014.

Titsias, M. Variational learning of inducing variables in sparse gaussian processes. In *Artificial Intelligence and Statistics*, pp. 567–574, 2009.

Ton, J.-F., Flaxman, S., Sejdinovic, D., and Bhatt, S. Spatial mapping with gaussian processes and nonstationary fourier features. *Spatial statistics*, 28:59–78, 2018.

Vaart, A. v. d. and Zanten, H. v. Information rates of nonparametric gaussian process methods. *Journal of Machine Learning Research*, 12(Jun):2095–2119, 2011.

van der Vaart, A. W., van Zanten, J. H., et al. Rates of contraction of posterior distributions based on gaussian process priors. *The Annals of Statistics*, 36(3):1435–1463, 2008.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Viña, A., Gitelson, A. A., Nguy-Robertson, A. L., and Peng, Y. Comparison of different vegetation indices for the remote assessment of green leaf area index of crops. *Remote Sensing of Environment*, 115(12):3468–3478, 2011.

Wang, K. A., Pleiss, G., Gardner, J. R., Tyree, S., Weinberger, K. Q., and Wilson, A. G. Exact gaussian processes on a million data points. *ArXiv*, abs/1903.08114, 2019.

Williams, C. K. and Seeger, M. Using the nyström method to speed up kernel machines. In *Advances in neural information processing systems*, pp. 682–688, 2001.

Yang, Q., Shi, L., and Lin, L. Plot-scale rice grain yield estimation using uav-based remotely sensed images via cnn with time-invariant deep features decomposition. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 7180–7183. IEEE, 2019.

You, J., Li, X., Low, M., Lobell, D., and Ermon, S. Deep gaussian process for crop yield prediction based on remote sensing data. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.