Who Supported Obama in 2012? Ecological Inference through Distribution Regression

Seth R. Flaxman
Machine Learning Department
and H. J. Heinz III College
Carnegie Mellon University
sflaxman@cs.cmu.edu

Yu-Xiang Wang
Machine Learning Department
Carnegie Mellon University
yuxiangw@cs.cmu.edu

Alexander J. Smola
Machine Learning Department
Carnegie Mellon University
and Marianas Labs
alex@smola.org

ABSTRACT

We present a new solution to the "ecological inference" problem, of learning individual-level associations from aggregate data. This problem has a long history and has attracted much attention, debate, claims that it is unsolvable, and purported solutions. Unlike other ecological inference techniques, our method makes use of unlabeled individual-level data by embedding the distribution over these predictors into a vector in Hilbert space. Our approach relies on recent learning theory results for distribution regression, using kernel embeddings of distributions. Our novel approach to distribution regression exploits the connection between Gaussian process regression and kernel ridge regression, giving us a coherent, Bayesian approach to learning and inference and a convenient way to include prior information in the form of a spatial covariance function. Our approach is highly scalable as it relies on FastFood, a randomized explicit feature representation for kernel embeddings. We apply our approach to the challenging political science problem of modeling the voting behavior of demographic groups based on aggregate voting data. We consider the 2012 US Presidential election, and ask: what was the probability that members of various demographic groups supported Barack Obama, and how did this vary spatially across the country? Our results match standard survey-based exit polling data for the small number of states for which it is available, and serve to fill in the large gaps in this data, at a much higher degree of granularity.

Keywords

Machine learning; supervised learning; kernel methods; Gaussian processes; distribution regression

1. INTRODUCTION

The name ecological inference refers to the idea of ecological correlations [28], that is correlations between variables observed for a group of individuals, as opposed to individual correlations, where the individuals are the unit of analysis. The ecological inference problem has much in common with the "modifiable areal unit problem" [20] and Simpson's paradox. Simply put, it is the problem of inferring individual correlations from ecological correlations. This challenge arises in computational advertising, healthcare data, opinion survey data, and population health data, because in each case for privacy or cost reasons, we are missing individual-level data, we have access to aggregate-level data, and we want to make individual-level predictions. One way to understand the reason it is called a "problem" is to consider a two-by-two contingency table, with unknown entries inside the table, and known marginals. As shown in the contingency table below, we might know that a certain electoral district's voting population is 43% men and 57% women and that in the last election, the outcome was 63% in favor of the Democratic candidate and 37% in favor of the Republican candidate. These percentages correspond to the numbers of individuals shown below: Is it possible to infer the

	Democrat	Republican	
Men	?	?	1,500
Women	?	?	2,000
	2,200	1,300	

joint and thus conditional probabilities, for example can we ask, what was the Democratic candidate's vote share among women voters? It is clear that only very loose bounds can be placed on these probabilities without any more information. Based on the fact that rows and columns must sum to their marginals, we know, e.g., that the number of Democrats who are men is between 0 and 1,500. These types of deterministic bounds have been around since the 1950's, under the name the method of bounds [4].

What if we are given a set of electoral districts, where for each we know the marginals of the two-by-two contingency table, but none of the inner entries? Then, thinking statistically, we might be tempted to run a regression, predicting the electoral outcomes based on the gender breakdowns of the districts. But this approach, formalized as Goodman's method [8] a few years after the method of bounds was proposed, can easily lead us astray—there is not even a guarantee that outcomes be bounded between 0 and 1, and it ignores potentially useful information provided by deterministic bounds.

We review related work in Section 2 and provide the necessary background on kernel embeddings of distributions, distribution regression, and GP regression in Section 3. We formalize the ecological inference problem in Section 4 and propose our method in Section 5. We apply it to the case of the 2012 US presidential election in Section 6, comparing our results to survey-based exit polls.

KDD'15, August 10-13, 2015, Sydney, NSW, Australia.

ACM 978-1-4503-3664-2/15/08 ...$15.00.
DOI: http://dx.doi.org/10.1145/2783258.2783300.

2. RELATED WORK

The ecological inference problem has a long history of solutions, counter-solutions, and it is often taught with a note of grave caution and stark warnings that ecological inference is to be avoided at all costs, usually in favor of individual-level surveys. As with Simpson's paradox, it should come as no surprise that correlations at one level of aggregation can and do flip signs at other levels of aggregation. But abandoning all attempts at ecological inference in favor of surveys is not feasible or appropriate in many circumstances—relevant respondents are no longer alive to answer historical questions of interest; subjects are reluctant to answer questions about sensitive topics like drug usage or cheating—meaning social scientists have been hard-pressed and even discouraged from studying many interesting and important questions. Ecological inference problems appear in demography, sociology, geography, and political science, and—as discussed in [13]—landmark legislation in the US such as the Voting Rights Act requires a solution to the ecological inference problem to understand racial voting patterns[1].

This problem has attracted a variety of approaches over the years as summarized in [13], which also proposes a Bayesian statistical modeling framework incorporating the method of bounds (thus uniting the deterministic and probabilistic approaches). [13] sparked a renewed interest in ecological inference, much of which is summarized in [14]. A parametric Bayesian approach to this setting was proposed in [12] and a semiparametric approach was proposed in [23].

Our method differs from existing methods in fours ways. First, it uses more information than is typically considered in a standard ecological regression setting: we assume that we have access to representative *unlabeled* individual-level data. In the voting example, this means having a sample of individual-level census records ("microdata") about each electoral district. Second, our method incorporates spatial variation. Spatial data is a common feature of ecological regressions (which, after all, usually have much to do with geography) but it is only very recently that ecological inference methods have begun to address spatial variation explicitly [14]. Third, while our method may be applied to the classic ecological inference problem of inferring individual level correlations from aggregate data, we propose that it is most well-suited to a related ecological problem, common in political science: inferring the unobserved behavior of subgroups based on the aggregate behavior of groups of which they are part. For our application, this means inferring the voting behavior of men and women separately by electoral district, given aggregate voting information by district. Finally, our work is nonparametric. Kernel embeddings are used to capture all moments of the probability distribution over covariates, and Gaussian process regression is used to non-parametrically model the dependence between predictors and labels.

A related line of work, termed "learning from label proportions" by some authors [24, 15, 30, 21], has the individual-level goal in mind, and aims to build a classifier for individual instances based only on group level label proportions. While in principle, this approach could be used in our setting, since we are only interested in subgroup level predictions the extra task of estimating individual level predictions is probably not worth the effort considering we are working with $n = 10$ million individuals.

Our method is based on recent advances in distribution regression [6, 33], which we generalize to address the ecological inference case. Previous work on distribution regression has relied on kernel ridge regression, but we use Gaussian process (GP) regression instead, thus enabling us to incorporate spatial variation, learn kernel hyperparameters, and provide posterior uncertainty intervals, all in a fully Bayesian setting. For scalability (our experiments use $n = 10$ million individuals), we use a randomized explicit feature representation ("FastFood") [16] rather than the kernel trick.

[1]Long-standing solutions have proved quite inadequate: in one court case involving the Voting Rights Act, a qualified expert testified, based on Goodman's method, that the percentage of blacks who were registered to vote in a certain electoral district exceeded 100% [13]. This evidently false claim was apparently made earnestly.

3. BACKGROUND

In this section we review kernel embeddings for probability distributions, distribution regression, FastFood, and Gaussian process (GP) regression.

3.1 Kernel embeddings of distributions

Kernel embeddings of distributions, e.g. [31, 32, 5], are a powerful class of reproducing kernel Hilbert space (RKHS) techniques that map joint, marginal and conditional probability distributions to vectors in a high (or infinite) dimensional feature space. Let $\phi : \mathbb{R}^n \to \mathcal{H}$. It has been shown that if a kernel map ϕ is universal/characteristic (e.g. a Gaussian RBF kernel), then for iid samples $x \sim X$, the mean embedding in feature space, denoted:

$$\mu_X = \mathbb{E}_{x\sim X}[\phi(x)] \tag{1}$$

completely characterizes the distribution in the sense that any two distributions with a difference in any moment will be mapped to a different point in the Hilbert space. This result has been used in a variety of kernel-based statistical tests, including tests of independence and two-sample tests [10]. It is a key feature of our method, because it will allow us to link aggregate labels to individual-level data without throwing out any information.

In this work, we use the simple empirical mean estimator for the kernel mean:

$$\widehat{\mu_X} = \frac{1}{N}\sum_j \phi(x^j) \tag{2}$$

It is shown in Smola et al. [31] that this plug-in estimator is a consistent, and it converges to μ_X with rate $O(\mathcal{R}_n(\mathcal{H}) + 1/\sqrt{n})$, where $\mathcal{R}_n(\mathcal{H})$ is the Rademacher complexity of the RKHS. As long as $\mathcal{R}_n(\mathcal{H}) = O(n^{-1/2})$ we have the (optimal) parametric rate. Recent work has focused on improving this estimator using James-Stein shrinkage [17].

3.2 Distribution regression

In this section, we formalize distribution regression, the task of learning a classifier or a regression function that maps probability distributions to labels. The problem is fundamentally challenging because we only observe the probability distributions through groups of samples from these distributions. Specifically, our dataset is structured as follows:

$$\left(\{x_1^j\}_{j=1}^{N_1}, y_1\right), \left(\{x_2^j\}_{j=1}^{N_2}, y_2\right), \ldots \left(\{x_n^j\}_{j=1}^{N_n}, y_n\right) \tag{3}$$

where group i has a single real-valued label y_i and N_i individual observations (e.g. demographic covariates for N_i individuals) denoted $x_i^j \in \mathbb{R}^d$.

To admit a theoretical analysis, it is assumed that the probability distributions themselves are drawn randomly from some unknown meta distribution of probability distributions. The intuition behind why distribution regression is possible is that if each group of samples are iid draws from a distribution which is itself an iid drawn from the meta distribution, then we will be able to learn.

Recently, this "two-stage sampled" structure was analyzed, showing that a ridge regression estimator is consistent [33] with polynomial rate of convergence for almost any meta-distribution of distributions that are sufficiently smooth. The basic approach is as follows: use the kernel mean estimator of Eq. (2) for each group separately to estimate:

$$\widehat{\mu_1} = \frac{1}{N_1}\sum_{j=1}^{N_1}\phi(x_1^j), \quad \ldots, \quad \widehat{\mu_n} = \frac{1}{N_n}\sum_{j=1}^{N_n}\phi(x_n^j) \tag{4}$$

Next, use kernel ridge regression [29] to learn a function f:

$$y = f(\widehat{\mu}) + \epsilon \tag{5}$$

where the objective is to minimize the L_2 loss subject to a "ridge" complexity penalty weighted by a positive constant λ:

$$\hat{f} = \arg\min_{f\in\mathcal{H}_f}\sum_i [y_i - f(\widehat{\mu_i})]^2 + \lambda\|f\|^2_{\mathcal{H}_f} \tag{6}$$

In [33] a variety of kernels for f corresponding to the Hilbert space $\mathcal{H}_f$ are considered. We follow the simplest choice of the linear kernel $k(\widehat{\mu_i}, \widehat{\mu_j}) = \langle\widehat{\mu_i}, \widehat{\mu_j}\rangle$, motivated by the fact that we are already working in Hilbert space over the μ_i. Following the standard derivation of kernel ridge regression [29], we can find the function f in closed form for a new test group μ_*:

$$f(\mu_*) = k^*(K + \lambda I)^{-1}[y_1, \ldots, y_n]^T \tag{7}$$

where $k^* = [\langle\widehat{\mu_1}, \mu_*\rangle, \ldots, \langle\widehat{\mu_n}, \mu_*\rangle]$ and $K_{ab} = \langle\widehat{\mu_a}, \widehat{\mu_b}\rangle$. In practice, it is hard to know whether the conditions under which the proofs in these papers hold are met. As a partial remedy, our Bayesian approach allows us to quantify the degree of uncertainty in our posterior predictions. Also, as shown in the experiments, a useful diagnostic is to measure the distance between training and test distributions.

3.3 FastFood for explicit kernel expansion

Naively implementing distribution regression using the kernel trick is not scalable in the setting we consider: to compute just one entry in K requires computing $K_{ab} = \langle\widehat{\mu_a}, \widehat{\mu_b}\rangle = \frac{1}{N_a N_b}\sum_{j_1 j_2} k(x_a^{j_1}, x_b^{j_2})$. This computation is $O(N^2)$ (where we assume for simplicity $N_i = N$, $\forall i$) so computing K is $O(n^2N^2)$. In our application, $N \approx 10^4$, so we need a much more scalable approach. Since we ultimately only need to work with the mean embeddings μ_i rather than the individual observations x_i^j, an explicit feature representation, even if it is very high-dimensional, will drastically reduce our computational costs.

We use an approximate kernel transformation called FastFood [16], which finds a d-dimensional approximation $\hat{\phi}(x) \in \mathbb{R}^d$ of $\phi(x)$ for every x. Here ϕ can be any radial basis function (RBF) kernel. Take Gaussian RBF kernel as an example, FastFood boils down to the following transformation due to Rahimi and Recht [25]:

$$\phi(x) = p^{-1/2}\exp(i[Vx])$$

where i is the imaginary unit of a complex number and V is a appropriately scaled $p \times d$ Gaussian random matrix with $p > d$. FastFood allows us to approximately compute Vx without explicitly construct V. In particular, FastFood transformation takes $V = [V_1^T, V_2^T, ..., V_{\lceil p/d\rceil}^T]^T$ and each square $d \times d$ matrix is given by:

$$V_j = \frac{1}{\sigma\sqrt{d}} SHG\Pi HB,$$

Here S, B, G are diagonal random matrices (nonnegative scaling, Rademacher and Gaussian respectively), Π is a random permutation, and H is the Walsh-Hadamard matrix. Every single one of these transformation can be computed in almost linear time. The whole transformation $\phi(x)$ can be therefore computed in $O(p\log d)$ time. This is orders of magnitude faster than random kitchen sinks [25] which costs $O(pd)$ per transformation or the kernel trick which needs to do an $O(N^3)$ inversion of a dense $N \times N$ kernel matrix.

It is shown in [16, Theorem 6] that for any x, x', $\langle\hat{\phi}(x), \hat{\phi}(x')\rangle$ converges to $\langle\phi(x), \phi(x')\rangle$ with rate $O(\frac{\log(2/\delta)}{p^{1/2}})$ where δ is the failure probability. This can be viewed as a Johnson-Lindenstrauss transformation of an infinite dimensional space to a finite dimensional Euclidean space while preserving the angles and distances in the original space. While this is not a uniform convergence bound as with the random features in [25], the exponential tail enables us to simultaneously guarantee an exponential number of kernel evaluations via the union bound. Not surprisingly, it has been empirically shown to be comparable in accuracy and to approximate the kernel transformation for all data points quite well. For simplicity, from here onwards we will overload notation and refer to $\phi(x) \in \mathbb{R}^p$ as our feature mapping which is understood to be approximated with FastFood.

3.4 Gaussian process regression

In this section, we briefly state the main results we need from Gaussian process regression [26], reviewing the well-known connection between the posterior mean in GP regression and the kernel ridge regression estimator of Eq. (7).

Given observations $(s_1, y_1), \ldots (s_n, y_n)$ a Gaussian process prior on a function f where our model is $y = f(s) + \epsilon$ is written:

$$f \sim \mathcal{GP}(0, k(s, s'))$$

with mean 0 and covariance function k. This implies that for a finite set of locations $X = \{s_1, \ldots, s_n\}$, the distribution of $\mathbf{f} = [f(s_1), \ldots, f(s_n)]^\top$ is multivariate Gaussian:

$$\mathbf{f} \sim \mathcal{N}(0, K) \tag{8}$$

where $K_{ij} = k(s_i, s_j)$. Notice that we have switched from a function $f(s)$ to a vector $\mathbf{f}$. This is because it is only formally correct to consider a probability distribution over the finite-dimensional vector $\mathbf{f}$, not over the infinite dimensional function $f(s)$. For a formal discussion see [35]. Conditional on the latent variable $\mathbf{f}$, we have a Gaussian observation model:

$$y_i|f(s_i) \sim \mathcal{N}(0, \sigma^2), \quad \forall i \tag{9}$$

for variance parameter σ^2 which can be thought of as measurement error (known as the "nugget" in geostatistics). For a fixed set of locations X, it is straightforward to sample $\mathbf{f}$ from its prior distribution in Eq. (8). Due to conjugacy, we can marginalize out $\mathbf{f}$ in closed form to find the distribution:

$$\mathbf{y} \sim \mathcal{N}(0, K + \sigma^2 I) \tag{10}$$

If we wish to make a prediction at a new location s^*, the standard predictive equations for GP regression [26], derived by conditioning a multivariate Gaussian distribution, tell us:

$$y^* \mid s^*, X, \mathbf{y} \sim \mathcal{N}(k^*(K+\sigma^2 I)^{-1}\mathbf{y}, k^{**} - k^*(K+\sigma^2 I)^{-1}k^{*\top}) \tag{11}$$

where $K_{ij} = k(s_i, s_j)$ and $k^* = [k(s_1, s^*) \dots k(s_n, s^*)]$ and $k^{**} = k(s^*, s^*)$. Thus we have a way of combining a prior over $\mathbf{f}$, parametrized by $k(s, s')$, with observed data to obtain a posterior distribution over a new prediction y^* at a new location s^*. This is a very powerful method, as it enables a fully Bayesian treatment of regression, a coherent approach to kernel learning through the marginal likelihood (for details see [26]), and posterior uncertainty intervals.

We can immediately see the connection between the kernel ridge regression estimator in Eq. (7) and the posterior mean of the GP in Eq. (11). (A superficial difference is that in Eq. (7) our predictors are $\widehat{\mu_i}$ while in Eq. (11) they are generic locations s_i, but this difference will go away in Section 5 when we propose using GP regression for distribution regression.) The predictive mean of GP regression is exactly equal to the kernel ridge regression estimator, with σ^2 corresponding to λ. In ridge regression, a larger penalty λ leads to a smoother fit (equivalently, less overfitting), while in GP regression a larger σ^2 favors a smoother GP posterior because it implies more measurement error. For a full discussion of the connections see [2, Sections 6.2.2-6.2.3].

4. ECOLOGICAL INFERENCE

In this section we state the ecological inference problem that we intend to solve. We use the motivating example of inferring Barack Obama's vote share by demographic subgroup (e.g. men versus women) in the 2012 US presidential election, without access to any individual-level labels. Vote totals by electoral precinct are publicly available, and these provide the labels in our problem. Predictors are in the form of demographic covariates about individuals (e.g. from a survey with individual level data like the census). The challenge is that the labels are aggregate, so it is impossible to know which candidate was selected by any particular individual. This explains the terminology: "ecological correlations" are correlations between variables which are only available as aggregates at the group level [28]

We use the same notation as in Section 3.2. Let $x_i^j \in \mathbb{R}^d$ be a vector of covariates for individual i in region j. Let w_i^j be survey weights[2]. Let y_i be labels in the form of two-dimensional vectors (k_i, n_i) where k_i is the number of votes received by Obama out of n_i total votes in region i. Then our dataset is:

$$\left(\{x_1^j\}_{j=1}^{N_1}, y_1\right), \left(\{x_2^j\}_{j=1}^{N_2}, y_2\right), \dots, \left(\{x_n^j\}_{j=1}^{N_n}, y_n\right) \tag{12}$$

We will typically have a rich set of covariates available, in addition to the demographic variables we are interested in stratifying on, so the x_i^j will be high-dimensional vectors denoting gender, age, income, education, etc.

Our task is to learn a function f from a demographic subgroup (which could be everyone) within region i to the probability that this demographic subgroup supported Obama, i.e. the number of votes this group gave Obama divided by the total number of votes in this group.

[2]Covariates usually come from a survey based on a random sample of individuals. Typically, surveys are reported with survey weights w_i^j for each individual to correct for oversampling and non-response, which must be taken into account for any valid inference (e.g. summary statistics, regression coefficients, standard errors, etc.).

5. OUR METHOD

In this section we propose our new ecological inference method. Our approach is illustrated in a schematic in Figure 1 and formally stated in Algorithm 1.

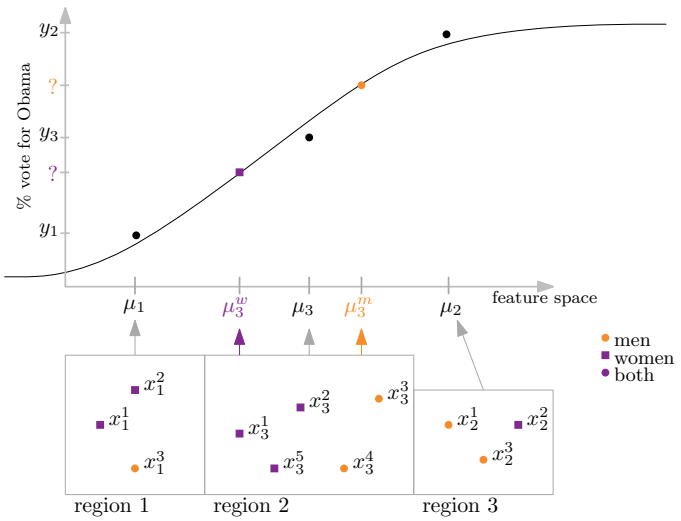


Figure 1: Illustration of our approach. Labels y_1, y_2 and y_3 are available at the group level giving Obama's vote share in regions 1, 2, and 3. Covariates are available at the individual level giving the demographic characteristics of a sample of individuals in regions 1, 2, and 3. We project the individuals from each group into feature space using a feature map $\phi(x)$ and take the mean by group to find high-dimensional vectors μ_1, μ_2 and μ_3, e.g. $\mu_1 = \frac{1}{3}(\phi(x_1^1) + \phi(x_1^2) + \phi(x_1^3))$. Now our problem is reduced to supervised learning, where we want to learn a function $f : \mu \to y$. Once we have learned f we make subgroup predictions for men and women in region 3 by calculating mean embeddings for the men $\mu_3^m = \frac{1}{2}(\phi(x_3^3) + \phi(x_3^4))$ and women $\mu_3^w = \frac{1}{3}(\phi(x_3^1) + \phi(x_3^2) + \phi(x_3^5))$ and then calculating $f(\mu_3^m)$ and $f(\mu_3^w)$. For a more rigorous description of our algorithm see Algorithm 1.

Recall the two-stage distribution regression approach introduced in Section 3.2. Our method has a similar approach. To begin, we use FastFood as introduced in Section 3.3 with an RBF kernel to produce an explicit feature map ϕ and calculate the mean embeddings[3], one for each region i, of Eq. (4) with survey weights:

$$\widehat{\mu_1} = \frac{\sum_j w_1^j \phi(x_1^j)}{\sum_j w_1^j}, \quad \dots, \quad \widehat{\mu_n} = \frac{\sum_j w_n^j \phi(x_n^j)}{\sum_j w_n^j} \tag{13}$$

[3] Distribution regression with explicit random features was previously considered in Oliva et al. [19] using Rahimi and Recht [25] to speed up an earlier distribution regression method based on kernel density estimation [22]. This approach has comparable statistical guarantees to distribution regression using RKHS-mean embeddings but inferior empirical performance [33]. As far as we are aware, using FastFood kernel mean embeddings for distribution regression is a novel approach.

Algorithm 1 Ecological inference algorithm

Input: $\left(\{(x_1^j, w_1^j)\}_{j=1}^{N_1}, s_1, y_1\right), \ldots, \left(\{(x_n^j, w_n^j)\}_{j=1}^{N_n}, s_n, y_n\right)$

1: **for** $i = 1 \ldots n$ **do**
2: Calculate $\widehat{\mu_i}$ using Eq. (13) with FastFood.
3: Calculate μ_i^m using Eq. (17) with FastFood.
4: **end for**
5: Learn hyperparameters $\hat{\theta} = (\sigma_x^2, \sigma_s^2, \ell)$ of the GP model specified by Eqs. (14)–(15) with observations y_i at locations $(\widehat{\mu_1}, s_1), \ldots, (\widehat{\mu_n}, s_n)$ using gradient descent and the Laplace approximation.
6: Make posterior predictions using $\hat{\theta}$ at locations $(\mu_1^m, s_1), \ldots, (\mu_n^m, s_n)$ using the Laplace approximation.

Output: Posterior means and variances for $y_1^m, \ldots, y_n^m$

Next, instead of kernel ridge regression, we use GP regression. Recall that unlike in distribution regression our labels y_i are given by vote counts (k_i, n_i). We use a Binomial likelihood as the observation model in GP regression (this is sometimes known as a logistic Gaussian process [27]). We transform each component of the latent real-valued vector $\mathbf{f}$ of Section 3.4 by the logistic link function $\sigma(\mathbf{f}) = \frac{1}{1+e^{-\mathbf{f}}}$ and we replace Eq. (9) with the following:

$$k_i | f(x_i) \sim \text{Binomial}(n_i, \sigma(f(x_i))) \tag{14}$$

where we use the formulation for the Binomial distribution of n_i trials and probability of success $\sigma(f(x_i))$. This is the generalized linear model (GLM) specification for binary data, combining a Binomial distribution with logistic link function [3, Ch. 7].

The predictors in our GP are the mean embeddings $\widehat{\mu_1}, \ldots, \widehat{\mu_n}$. We also include spatial information in the form of 2-dimensional spatial coordinates s_i giving the centroid of region i. Putting these predictors together we adopt an additive covariance structure:

$$\mathbf{f} \sim \mathcal{GP}(0, \sigma_x^2 \langle \widehat{\mu_i}, \widehat{\mu_j} \rangle + k_s(s_i, s_j)) \tag{15}$$

Where we have used a linear kernel between mean embeddings weighted by a variance parameter σ_x^2. Since the mean embeddings are already in feature space using the FastFood approximation to the RBF kernel, we are approximately using the RBF kernel. For the spatial coordinates we use the Matérn covariance function which is a popular choice in spatial statistics [11], with $\nu = 3/2$, length-scale ℓ and variance parameter σ_s^2:

$$k(s, s') = \sigma_s^2 \left(1 + \frac{\|s - s'\|\sqrt{3}}{\ell}\right) \exp\left(-\frac{\|s - s'\|\sqrt{3}}{\ell}\right) \tag{16}$$

By adding together the linear kernel between mean embeddings and the spatial covariance function, we allow for a smoothly varying surface over space and demographics. The intuition is that this additive covariance encourages predictions for regions which are nearby in space and have similar demographic compositions to be similar; predictions for regions which are far away or have different demographics are allowed to be less similar. GP regression with a spatial covariance function is equivalent to the spatial statistics technique of kriging—we are effectively smoothly interpolating y values over a very high dimensional space of predictors. Another way to think about additivity is that we are accounting for a spatially autocorrelated error structure in the predictions we get from covariates alone. (We also considered a multiplicative structure, which had slightly worse performance.)

Eq.s (14)-(15) complete our hierarchical model specification. For non-Gaussian observation models like Eq. (14), the posterior prediction in Eq. (11) is no longer available in closed form due to non-conjugacy. We follow the standard approach for GP classification and logistic Gaussian processes and use the Laplace approximation [36, 27]. The Laplace approximation gives an approximate posterior distribution for $\mathbf{f}$, from which we can calculate a posterior distribution over the k_i of Eq. (14) as explained in detail in [26, Section 3.4.2]. The Laplace approximation also allows us to calculate the marginal likelihood, which is the probability of the observed data, integrating out $\mathbf{f}$. To learn σ_x^2, σ_s^2, and ℓ, we use gradient ascent to maximize the log marginal likelihood.

Once we have learned the best set of hyperparameters for our model we can make predictions for any demographic subgroup of interest. To predict the fraction of men who voted for Obama, we create new mean embedding vectors by gender and region, modifying Eq. (13):

$$\widehat{\mu_i^m} = \frac{\sum_{j^m} w_1^j \phi(x_1^j)}{\sum_{j^m} w_1^j}, \quad \forall i \tag{17}$$

where j^m are the indices of the observations of men in region i and $\widehat{\mu_i^m}$ is the mean embedding of the covariates for the men in region i. We then make posterior predictions using the Laplace approximation as above at these new gender-region predictors. Notice that for a new μ^* this requires calculating $k^* = [k_{1*}, k_{2*}, \ldots, k_{n*}]$ of Eq. (11) where $k_{i*} = \sigma_x^2 \langle \widehat{\mu_i}, \mu_* \rangle + k_s(s_i, s_*)$ using Eq. (15). Thus new predictions will be similar to existing predictions in regions with similar covariates and they will be similar to existing predictions at the same (and nearby) locations.

Our algorithm is stated in Algorithm 1. We now analyze its complexity. Lines 2–3 are calculated by streaming through the data for individuals. For each individual, calculating the FastFood feature transformation $\phi(x_i^j)$ takes $\mathcal{O}(p \log d)$ where $x_i^j \in \mathbb{R}^d$ and $\phi(x_i^j) \in \mathbb{R}^p$. To save memory, there's no need to store each $\phi(x_i^j)$. We simply update the weighted average $\widehat{\mu_i}$ by adding $w_j^i \phi(x_i^j)$ to it. Notice that the demographic subgroup considered in line 3 is simply a subset of the observations calculated in line 2, so there is no added cost to calculate the μ_i^m or indeed a set of $\mu_i^{m_1}, \ldots, \mu_i^{m_q}$ for q different demographic subgroups of interest. Overall, if we have N individuals the for loop takes time $\mathcal{O}(Np \log d)$. Usually $p \ll N$ and $d \ll N$ so this is practically linear and trivially parallelizable.

On line 5 to learn the hyperparameters in the GP regression requires calculations involving the covariance matrix $K \in \mathbb{R}^{n \times n}$. Each entry in K requires computing a dot product $\langle \widehat{\mu_i}, \widehat{\mu_j} \rangle$ which takes $\mathcal{O}(p)$ and it requires computing the Matérn kernel for the spatial locations, which is a fast arithmetic calculation. Once we have K, the Laplace approximation is usually implemented with Cholesky decompositions for numerical reasons. The runtime of computing the marginal likelihood and relevant gradients is $\mathcal{O}(n^3)$ [26], and gradient ascent usually takes less than a hundred steps to converge. Posterior predictions on line 6 require calculating $k^* \in \mathbb{R}^{1 \times n}$ for each μ_i^m so this is $O(n^2)$. Reusing the Cholesky decompositions above means predictions can be made in $\mathcal{O}(n^2)$. GP regression requires $\mathcal{O}(n^2)$ storage.

Overall, we expect $n \ll N$, so our algorithm is practically $\mathcal{O}(N)$, with little extra computational cost arising from the GP regression as compared to the work of streaming through all the observations. The N observations do not need to be stored in memory, so the overall memory complexity is only $\mathcal{O}(n^2)$.

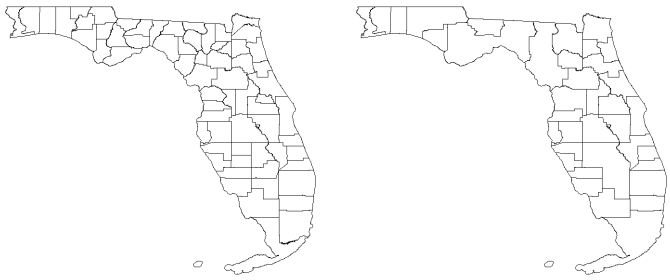
6. EXPERIMENTS

In this section, we describe our experimental evaluation, using data from the 2012 US Presidential election, and compare our results to survey-based exit polls, which are only available for the 18 states for which large enough samples were obtained. Our method enables us to fill in the full picture, with much finer-grained spatial estimation and results for a much richer variety of demographic variables. This demonstration shows the applicability of our new method to a large body of political science literature (see, e.g. [7]) on voting patterns by demographics and geography. Because voting behavior is unobservable and due to the ecological inference problem, previous work has been mostly based on exit polls or opinion polls.

We obtained vote totals for the 2012 US Presidential Election at the county level[4]. Most voters chose to either reelect President Barack Obama or vote for the Republican party candidate, Mitt Romney. A small fraction of voters ($< 2\%$ across the country) chose a third party candidate. Separately, we obtained data from the US Census, specifically the 2006-2010 American Community Survey's Public Use Microdata Sample (PUMS). The American Community Survey is an ongoing survey that supplements the decennial US census and provides demographically representatives individual-level observations. PUMS data is coded by public use microdata areas (PUMAs), contiguous geographic regions of at least 100,000 people, nested within states. We used the 5-year PUMS file (rather than a 1-year or 3-year sample) because it contains a larger sample and thus there is less censoring for privacy reasons. To merge the PUMS data with the 2012 election results, we created a mapping between counties and PUMAs[5], merging individual-level census data and aggregating vote totals as necessary to create larger geographic regions for which the census data and electoral data coincided. The mapping between PUMAs and counties is many-to-many, so we were effectively finding the connected components. Since counties and PUMAs do not cross state borders, none of the geographic regions we created cross state borders. An example is shown in Figure 2.

In total, we ended up with 837 geographic regions ranging from Orleans Parish in New Orleans, which voted 91% for Barack Obama to Davis County, a suburb of Salt Lake City, Utah which voted 84% for Mitt Romney. For the census data, we excluded individuals under the age of 18 (voting age in the US) and non-citizens (only citizens can vote in presidential elections). There were a total of 10,787,907 individual-level observations, or in other words, almost 11 million people included in the survey. The mean number of people per geographic region was 12,812 with standard deviation 21,939.

There were 223 variables in the census data, including both categorical variables such as race, occupation, and educational attainment and real valued variables such as income in past 12 months (in dollars) and travel time to work (in minutes). We divided the real-valued variables by their standard deviation to put them all on the same scale. For the categorical variables with D categories, we converted them into D dimensional 0/1 indicator variables, i.e. for the variable "when last worked" with categories 1 = "within the past 12 months," 2 = "1-5 years ago," and 3 = "over 5 years ago or never worked" we mapped 1 to $[1\ 0\ 0]^T$, 2 to [0 1 0] and 3 to [0 0 1].



(a) **(b)**

Figure 2: Election outcomes were available for the 67 counties in Florida shown in (a). Demographic data from the American Community Survey was available for 127 public use microdata areas (PUMAs) in Florida, which sometimes overlapped parts of multiple counties and sometimes contained multiple counties. We merged counties and PUMAs as described in text to create a set of disjoint regions with the result of 37 electoral regions as shown in (b).

Putting together the indicator variables and real-valued variables, we ended up with 3,251 variables total. For every single individual-level observation, we used FastFood with an RBF kernel to generate a 4,096-dimensional feature representation. Using Eq. (13) we calculated the weighted mean embedding for each region. The result was a set of 837 vectors which were 4,096-dimensional.

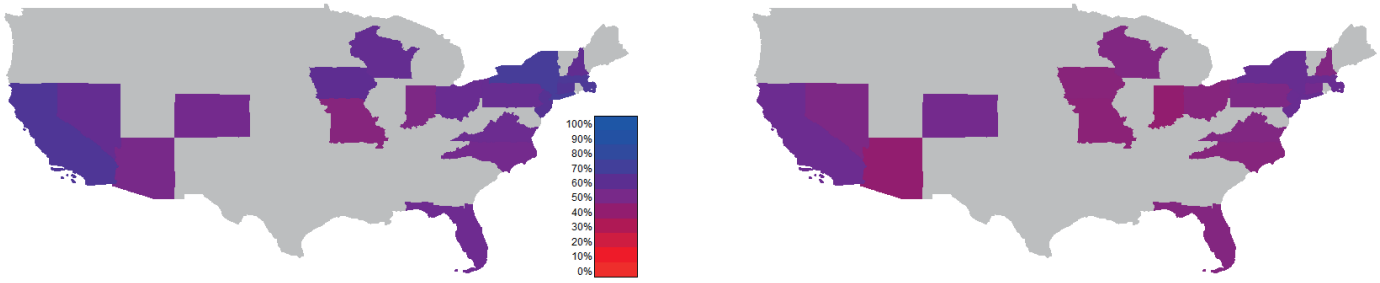
We treated the vote totals for Obama and Romney as is, discarded the remaining third party votes as the exit polls we use for validation did not report third party votes. Thus for each region, we had a positive integer valued 2-dimensional label giving the number of votes for Obama and the total number of votes.

We focused on the ecological inference problem of predicting Obama's vote share by the following demographic groups: women, men, income $\leq$ US$50,000 per year, income between $50,000 and $100,000 per year, income $\geq$ 100,000 per year, ages 18-29, 30-44, 45-64, and 64 plus. For each region, we used the strategy outlined above, restricting our census sample to only those observations matching the subgroup of interest and creating new mean embedding predictors as in Eq. (17), μ_i^{subgroup}. We made predictions for each region-demographic pair. Note that we have made our task harder than necessary to demonstrate our method; we could have trained our model using the exit polling data, where available, and we would certainly recommend practitioners use all available data to get the best possible estimates.

All of our models were fit using the GPstuff package with scaled conjugate gradient optimization and the Laplace approximation [34]. Since $n \ll N$, the time required to fit the GP model and make predictions is much less than the time

[4]https://github.com/huffpostdata/election-2012-results

[5]using the PUMA 2000 codes and the tool at http://mcdc.missouri.edu/websas/geocorr12.html



(a) Exit poll results for women

(b) Exit poll results for men

(c) Ecological regression results for women

(d) Ecological regression results for men

Figure 3: Support for Obama among women (a) and men (b) in the 18 states for which exit polling was done; due to cost, no representative data was collected for the majority of states or for regions smaller than states. Support for Obama among women (c) and men (d) in 837 different regions as inferred using our ecological regression method.

required to preprocess the data to create the mean embeddings at the beginning of Algorithm 1.

7. RESULTS

We learned the following hyperparameters for our GP: $\sigma_s^2 = 0.18$, $\ell = 7.92$, and $\sigma_x^2 = 4.56$. The σ^2 parameters can be roughly interpreted as the "fraction of variance explained" so the fact that σ_x^2 is much larger than σ_s^2 means that the demographic covariates encoded in the mean embedding are much more important to the model than the spatial coordinates. The length-scale for the Matérn kernel is a little more than half the median distance between locations, which indicates that it is performing a reasonable degree of smoothing. We used 10-fold crossvalidation to evaluate our model and ensure that it was not overfitting, an important consideration as generalization performance is critical. The root mean squared error of the model was 2.5 and the mean log predictive density was -1.9. Predictive density is a useful measure because it takes posterior uncertainty intervals into account. For comparison, predicting the national average of Obama receiving 51.1% of the vote in every location has a root mean squared error of 8.3. As a sensitivity analysis, we also considered a multiplicative model, for which the performance was comparable.

To validate our models, we compared to the 2012 exit polls, conducted by Edison Research for a consortium of news organizations. National results were based on interviews with voters in 350 randomly chosen precincts, and state results in 18 states were based on interviews in 11 to 50 random precincts. In these interviews, conducted as voters left polling stations, voters were asked who they voted for and a variety of demographic questions about themselves. Bias due to factors such as unrepresentativeness of the sampled precincts and inadequate coverage of early or absentee voters could be an issue [1]. The national results had a margin of error (corresponding to a 95% uncertainty interval) of 4 percentage points[6] and the state results had a margin of error of between 4 and 5 percentage points [18]. For comparing to the 18 state-level exit polls, we aggregated our geographic regions by state, weighting by subgroup population.

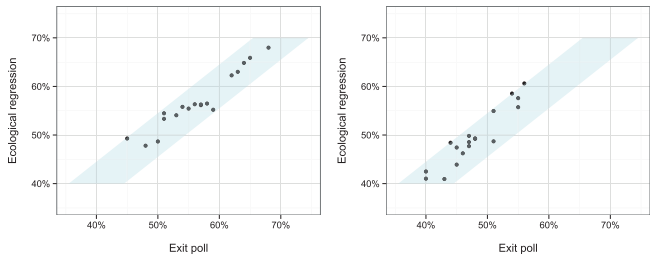
As a preview of our results by gender, income, and age, and to get an idea of the power of our method, Figure 3 shows four maps visualizing Obama's support among women and men. In Figures 3a–3b, we show the results from the exit polls, at the state level, for only 18 states. In Figures 3c–3d we fill in the missing picture, providing estimates for 837 different regions. We compare to competing methods below for national-level gender estimates. In the supplementary materials, we consider the non-binary demographic covariates age and income and the case of regional-level estimates, which present a difficulties for the competing methods.

7.1 Gender

Voting by gender is shown in Figure 4, where we compare our results to the exit poll results. The fit is quite good, with correlations equal to 0.96 for men and 0.94 for women. The inference that we are most interested in is the gender

[6]This presumably corresponds to a sample size of only $n = 600$ individuals, since the usual margin of error reported by news organizations is $1.96\sqrt{\frac{.5^2}{n-1}}$

gap—i.e. how much larger was Obama's vote share among women than among men? In Figure 5 we show a histogram of the gender gap by geographic region.



(a) Women **(b) Men**

Figure 4: Our model's ecological predictions (y-axis) of the probability of voting for Obama by gender compared to estimates obtained from an exit poll (x-axis). The blue line shows a 95% uncertainty interval around the 45° line, corresponding to uncertainty due to exit poll's margin of error of 4 percentage points.

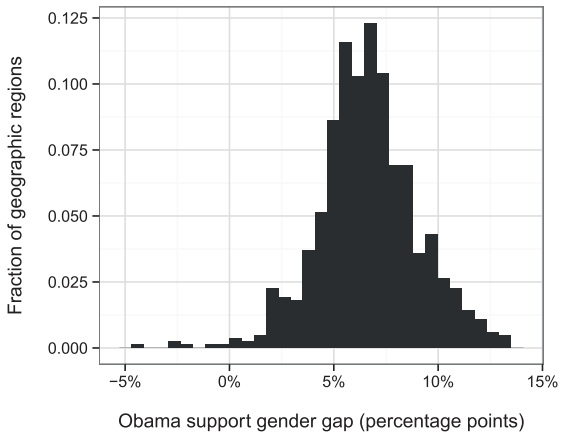


Figure 5: Gender gap is calculated as Obama support among women minus Obama support among men.

Taking a weighted average nationally, we can compare to existing methods and the ground truth. Our method's estimates exactly matched the ground truth from national-level exit polls: 48% support for Obama among men, 55% among women. An unkernelized version of our method estimated 50% among men, 53% among women. King's method estimated 29% among men, 72% among women, and Goodman's method estimated the impossible values of -75% among men, 158% among women.

7.2 Income

Voting by income is shown in Figure 6, where we compare our results to the exit poll results. In this plot, we have included both 95% uncertainty intervals and indicated the 95% margin of error from the exit poll. For low incomes ($\leq$ \$50,000) the correlation is 0.85, for medium incomes (between \$50,000 and \$100,000) the correlation is 0.90, and for high incomes ($\geq$ \$100,000) the correlation is 0.67. Compared to our gender predictions, it is clear that our model is not performing as well in terms of its mean predictions. On the other hand, it is clear that our model's uncertainty intervals are doing what they are designed to do: the large uncertainties in the posterior predictions for the high income group accurately reflect how much we should believe our posterior (recall that in the Bayesian paradigm these are "credibility intervals" rather than frequentist confidence intervals). To explore the reasons our predictions are less accurate, we considered the assumptions underlying distribution regression. In the two-stage analysis of [33], distributions are drawn from a meta distribution. When we make a prediction for a test distribution, if the test distribution is in a low density region of the meta distribution then we should not expect a reliable prediction. To test this assumption, we calculated k_* in Eq. (11) separately for each observation for low, medium, and high incomes as k_* provides a measure of the similarity between a test distribution and the distributions used to fit the model. The distribution of the values of k_* are shown in Figure 8, where they are compared to the entries in K, i.e. the "in-sample" regions for which we know the labels. While the distributions for low and medium income are quite close to the overall distribution, the distribution for high income is quite far. This is a useful diagnostic for distribution regression in general and ecological inference in particular. It might be possible to correct for this type of bias, as in the covariate shift literature [9].

7.3 Age

Voting by age is shown in Figure 7. The correlations are 0.60 (ages 18-29), 0.90 (30-44), 0.92 (45-64), and 0.90 (65 years or older). We do not include posterior uncertainty intervals for clarity, but as in the previous section, these seem to be properly calibrated: the average variance is 0.10 for ages 18-29, 0.06 for ages 30-44, 0.05 for ages 45-64, and 0.13 for ages 64 years or older.

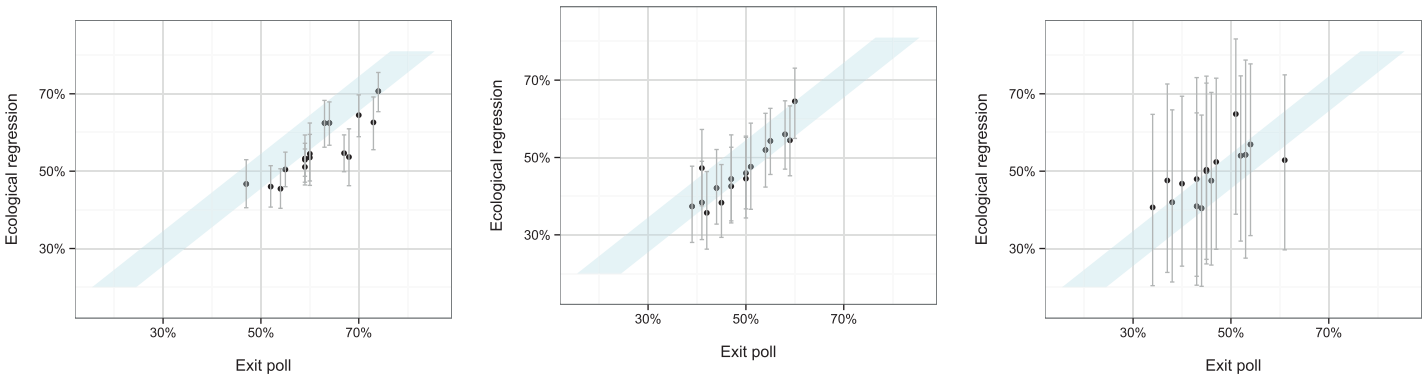
In Table 1, we compare our national estimates to estimates from the nationally representative exit poll. With the exception of the youngest age group, where we significantly overestimate Obama's support, our predictions are quite accurate, with our posterior predictions matching the exit polls to within just a few percentage points. Our results are weighted based on the percent of the population, rather than the percent of the voting age population. For example, 22% of residents in the US are aged 18-29, but only 19% of voters (according to the exit polls) were aged 18-29. This means that our mean embedding vectors are slightly biased, an issue that we intend to address in future work.

Age group	Ecological inference	% of residents	Exit poll [95% UI]	% of voters
18-29	70	22	60 [56, 64]	19
30-44	52	24	52 [48, 56]	27
45-64	45	35	47 [43, 51]	38
65+	43	18	44 [40, 48]	16

Table 1: For the US, we compare our ecological inference to nationally representative exit polls.

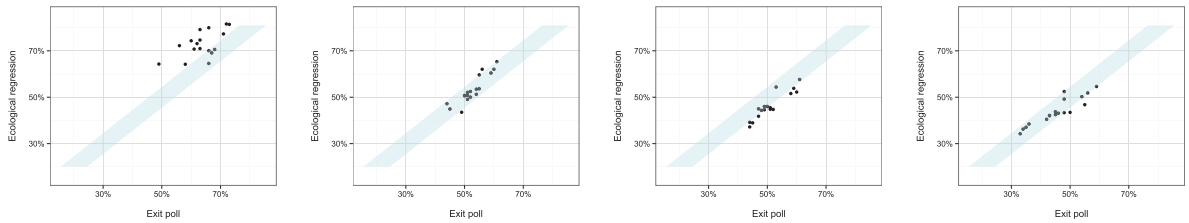
8. CONCLUSION

In this paper, we developed a new method to address the long-standing ecological inference problem. Our method



(a) Income $\leq$ \$50,000 (b) Income between \$50,000 and \$100,000 (c) Income $\geq$ \$100,000

Figure 6: Our model's predictions of Obama vote share by income compared exit polling.



(a) 18-29 year olds (b) 30-44 year olds (c) 45-64 year olds (d) 65 years or older

Figure 7: Our model's predictions of Obama vote share by age compared exit polling.

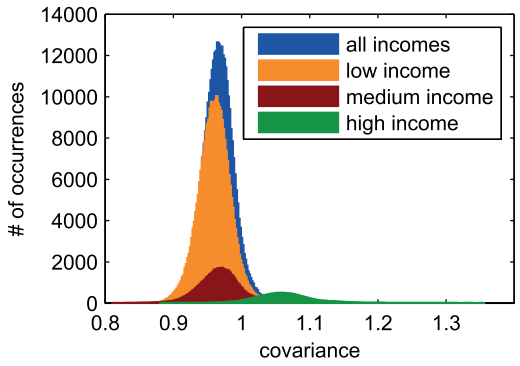


Figure 8: We calculated the vector k_* of Eq. (11) by income (low, medium, high) for each region and compared the distribution of the values of k_* to the distribution of the values of K ("all incomes").

makes use of information often left unused in standard ecological regression, that of unlabeled, individual-level data. We formulated a novel and scalable Gaussian process distribution regression method which naturally incorporates spatial information and enables Bayesian inference. Our model generated posterior predictions and uncertainty intervals and where the predictions were less accurate, the uncertainty intervals were larger.

As far as we are aware, distribution regression has not previously had a Bayesian treatment. Our approach also highlights the potential for GLM approaches to distribution regression. Our Gaussian process regression framework could be immediately extended to continuous, categorical, or multivariate output settings and to including other structure in the input space, such as graph or temporal constraints.

Our new method could be used in to trying to answer a variety of social science and public policy questions, especially to answer questions for which carrying out population-representative surveys is impossible or in settings in which the goal is to combine together two different group-level surveys. We used our method in an important political science setting, that of understanding voting patterns by demographic group. We were thus able to move towards filling an important gap in the political science literature about the 2012 US presidential election due to the lack of representative exit polls covering all 50 US states. Our model's predictions were quite accurate, despite the fact that we did not actually use all of the information at our disposal.

9. ACKNOWLEDGEMENTS

A.S. thanks Talia Borodin for an inspiring discussion on the subject of electoral statistics. The research was partially supported by the Singapore NRF under its International Research Centre at Singapore Funding Initiative, and NSF grants BCS-0941518 and IIS-0953330.

10. REFERENCES

[1] M. A. Barreto, F. Guerra, M. Marks, S. A. Nuño, and N. D. Woods. Controversies in exit polling:

Implementing a racially stratified homogenous precinct approach. *PS: Political Science and Politics*, 39(3):pp. 477–483, 2006. ISSN 10490965.

[2] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[3] A. Dobson. An introduction to generalized linear models. *Chapman & Hall texts in statistical science*, 2002.

[4] O. B. Duncan and D. B. An alternative to ecological correlation. *American Sociological Review*, 16:665–666, 1953.

[5] K. Fukumizu, L. Song, and A. Gretton. Kernel bayes' rule. In *NIPS*, pages 1737–1745, 2011.

[6] T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola. Multi-instance kernels. In *ICML*, volume 2, pages 179–186, 2002.

[7] A. Gelman, D. Park, B. Shor, J. Bafumi, and J. Cortina. *Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do*. Princeton University Press, Aug. 2008. ISBN 069113927X.

[8] L. A. Goodman. Some alternatives to ecological correlation. *American Journal of Sociology*, pages 610–625, 1959.

[9] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009.

[10] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *JMLR*, 13:723–773, 2012.

[11] M. S. Handcock and M. L. Stein. A bayesian analysis of kriging. *Technometrics*, 35(4):403–410, 1993.

[12] C. Jackson, N. Best, and S. Richardson. Improving ecological inference using individual-level data. *Statistics in medicine*, 25(12):2136–2159, 2006.

[13] G. King. *A Solution to the Ecological Inference Problem*. Princeton University Press, Mar. 1997.

[14] G. King, M. A. Tanner, and O. Rosen. *Ecological inference: New methodological strategies*. Cambridge University Press, 2004.

[15] H. Kueck and N. de Freitas. Learning about individuals from group statistics. In *21st Uncertainty in Artificial Intelligence (UAI)*, pages 332–339, 2005.

[16] Q. Le, T. Sarlos, and A. Smola. Fastfood: approximating kernel expansions in loglinear time. In *Proceedings of the international conference on machine learning*, 2013.

[17] K. Muandet, K. Fukumizu, B. Sriperumbudur, A. Gretton, and B. Schoelkopf. Kernel mean estimation and stein effect. In *Proceedings of The 31st International Conference on Machine Learning*, pages 10–18, 2014.

[18] New York Times. President exit polls - election 2012. `http://elections.nytimes.com/2012/results/president/exit-polls`, 2012. Accessed: 16 February 2015.

[19] J. B. Oliva, W. Neiswanger, B. Póczos, J. G. Schneider, and E. P. Xing. Fast distribution to real regression. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014, Reykjavik, Iceland, April 22-25, 2014*, volume 33 of *JMLR Proceedings*, pages 706–714. JMLR.org, 2014.

[20] S. Openshaw. Ecological fallacies and the analysis of areal census data. *Environment and Planning A*, 16 (1):17–31, 1984.

[21] G. Patrini, R. Nock, T. Caetano, and P. Rivera. (almost) no label no cry. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *NIPS 27*, pages 190–198. Curran Associates, Inc., 2014.

[22] B. Poczos, A. Singh, A. Rinaldo, and L. Wasserman. Distribution-free distribution regression. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 507–515, 2013.

[23] R. L. Prentice and L. Sheppard. Aggregate data studies of disease risk factors. *Biometrika*, 82(1): 113–125, 1995.

[24] N. Quadrianto, A. J. Smola, T. S. Caetano, and Q. V. Le. Estimating labels from label proportions. *JMLR*, 10:2349–2374, 2009.

[25] A. Rahimi and B. Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in neural information processing systems*, pages 1313–1320, 2008.

[26] C. E. Rasmussen and C. K. Williams. Gaussian processes for machine learning, 2006.

[27] J. Riihimäki and A. Vehtari. Laplace approximation for logistic gaussian process density estimation and regression. *Bayesian Analysis*, 9(2):425–448, 2014.

[28] W. S. Robinson. Ecological correlations and the behavior of individuals. *American Sociological Review*, 15:351–57, 1950.

[29] C. Saunders, A. Gammerman, and V. Vovk. Ridge regression learning algorithm in dual variables. In *(ICML-1998) Proceedings of the 15th International Conference on Machine Learning*, pages 515–521. Morgan Kaufmann, 1998.

[30] D. R. Sheldon and T. G. Dietterich. Collective graphical models. In *NIPS*, pages 1161–1169, 2011.

[31] A. Smola, A. Gretton, L. Song, and B. Schölkopf. A hilbert space embedding for distributions. In *Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

[32] L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968. ACM, 2009.

[33] Z. Szabo, A. Gretton, B. Poczos, and B. Sriperumbudur. Two-stage Sampled Learning Theory on Distributions. *Artificial Intelligence and Statistics (AISTATS)*, Feb. 2015.

[34] J. Vanhatalo, J. Riihimäki, J. Hartikainen, P. Jylänki, V. Tolvanen, and A. Vehtari. Gpstuff: Bayesian modeling with gaussian processes. *JMLR*, 14(1): 1175–1179, 2013.

[35] G. Wahba. *Spline models for observational data*, volume 59. Siam, 1990.

[36] C. K. Williams and D. Barber. Bayesian classification with gaussian processes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(12): 1342–1351, 1998.