Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing

Crowdsourcing Spatial Phenomena Using Trust-Based Heteroskedastic Gaussian Processes

Matteo Venanzi, Alex Rogers, Nicholas R. Jennings
University of Southampton
Southampton, UK
{mv1g10, acr, nrj}@ecs.soton.ac.uk

Abstract

Many crowdsourcing applications require spatial data modelling to make sense of location-based observations provided by multiple users. In this context, we propose a new spatial function modelling approach to address the problem of fusing multiple spatial observations reported by possibly untrustworthy users in the domains of participatory sensing and crowdsourcing applications. Specifically, we use a heteroskedastic Gaussian process model to incorporate user trust modelling into Bayesian spatial regression. In particular, by training the model with the reports gathered from the crowd, we are able to estimate the spatial function at any location of interest and also learn the level of trustworthiness of each user. We show that our method outperforms other standard homoskedastic and heteroskedastic Gaussian processes by up to 23% on a crowdsourced radiation dataset collected during the 2011 Fukushima earthquake in Japan. We also show that our method is able to improve the quality of spatial predictions on synthetic data by up to 70% and is robust in settings of up to 30% presence of untrustworthy users within the crowd.

Introduction

Participatory sensing is the paradigm of harnessing the power of ordinary people who voluntarily provide environmental readings using readily available sensor devices, such as smart phones or tablets. This paradigm has been successfully applied to crowdsourcing spatial data in various domains, including tracking contagious diseases (Sadilek, Kautz, and Silenzio 2012), monitoring traffic flows (Horvitz et al. 2012) and measuring nuclear radioactivity for environmental monitoring (Gertz and Di Justo 2012). In particular, the smart devices owned by the users are provided with a number of sensors such as microphone, camera and GPS sensor which enable them to report geo-tagged information contents. This rapid and inexpensive information gathering now provides an unprecedented amount of data that is useful to solve extremely important problems such as highly decentralised information gathering tasks in the domains mentioned above. However, one of the main obstacles to make use of such information is *data trustworthiness* which relates to the range of accuracy of the users in reporting their observations. In general, crowd generated content can be untrustworthy due to several dimensions of inaccuracy of humans as observers such as the errors of their sensor devices or the malicious behaviour of some users in reporting information strategically (Hall and Jordan 2010). Therefore, the task of aggregating the reports into a single estimate is difficult to achieve in practice. In particular, the computation of reliable aggregations of spatial data reported by untrustworthy users is a key challenge in crowdsourcing domains.

The challenge of merging untrustworthy information has started to be addressed within a number of AI communities. However, most of this work has focused on information fusion for crowdsourced classification and image labelling tasks. In these settings, the reports are typically represented as noisy samples of the fixed quantity observed by the crowd, i.e. the true object class or the image label. Then, the reports are fused using simple majority voting (Bachrach et al. 2012) or iterative learning methods (Reece et al. 2009), or using statistical models to infer both the accuracy of the users and the true answer to the task from the crowd responses (Dawid and Skene 1979; Raykar et al. 2010; Kamar, Hacker, and Horvitz 2012). However, these classification methods are unsuitable for dealing with regression problems involving spatial data since the spatial correlation within the report set introduces dependencies between the observed value and the observer's location. Therefore, the fusion of the reports must be derived as the continuous function estimating the crowdsourced spatial phenomenon which requires different inference approaches from the ones above.

To address these shortcomings, we develop a method for fusing crowdsourced spatial data in the setting where users have different unknown levels of trustworthiness. Our method builds upon the heteroskedastic Gaussian process (HGP) which is a powerful non-parametric learning model providing a flexible Bayesian inference framework for spatial regression (Rasmussen and Williams 2006). These qualities make such a model attractive to be employed for merging data also in crowdsourcing settings. Specifically, we develop a new method for aggregating crowdsourced spatial *estimates* where the reports consist of pairs of measurements and precisions. This setting is relevant to the large class of crowdsourcing problems where numerical values of the uncertainty about each observation is provided by the users as part of their reports. For example, such reported uncertainties may refer to the precision of a sensor, the variance of some repeated measurement, or the confidence level estimated through self-appraisal by the user.

In our HGP model, we introduce a set of trustworthiness hyperparameters to characterise the different users' reliabilities. We use the trust hyperparameters to uncertainty scaling parameters which provide the model with the ability to flexibly increase the noise around subsets of reports associated with untrustworthy users. Then, by training the model with the reports gathered from the crowd, we are able to estimate the underlying spatial function and also learn the individual user's trustworthiness. We show that our method is more accurate than other standard GP and HGP approaches with an extensive experimental evaluation on both synthetic and real-word data.

Thus, this paper makes the following contributions to the state of the art:

- We propose a trust-based HGP model which combines the HGP with a user trust model to be able to aggregate location-dependent crowdsourced observations while learning the individual user trustworthiness.
- We show that our method significantly improves the quality of the predictions of other GP and HGP methods in an application of crowdsourced radiation monitoring using real-world data from the 2011 Fukushima nuclear disaster. In particular, our method outperforms the benchmarks by up to 23%. We also provide an in-depth analysis of the performance using synthetic data showing that our method is robust in settings with up to 30% untrustworthy users and improves the predictions of up to 70%.

In the remainder of the paper, we first discuss the rest of the related work from community sensing and information fusion in spatial crowdsourcing. Then we describe our model and its inference process. Finally, we discuss our experimental results and conclude.

Related Work

Prior work on community sensing addresses the problem of reliably merging spatial information provided by multiple users in various applications. Krause et al. (2008) discuss optimal policies for the online integration of sensor information in community sensing applied to traffic monitoring data. Their approach focuses on modelling the online information acquisition process aiming to maximise the utility of the acquired information while taking into account the limited resources and system constraints. In the same domain, Herring et al. (2010) applies logistic regression techniques to estimate the congestion state of the roads from GPS reports. However, both of these approaches do not address the question of how to deal with untrustworthy reports in the data fusion process which is the focus of this work.

Faulkner et al. (2011) designed a system for decentralised detection of earthquakes using cell phone accelerometer data. In this setting, smart phones provide sensor readings from their accelerometers and compute the probability of an earthquake using a hierarchical hypothesis testing approach. Then, a decentralised decision-theoretic framework is used to merge local sensor decisions into a final earthquake prediction. However, their approach is only applicable to binary classification problems, e.g. earthquake events, whereas we focus on spatial regression problems with a continuous space of decision variables.

In a more comparable setting, Groot, Birlutiu, and Heskes (2011) applied the standard GP model to regression problems with multiple inaccurate annotators in object labelling tasks. In their model, the accuracy of each object label is taken as the aggregation of the accuracies of its annotators. Then, the individual object accuracies are incorporated in the GP as latent hyperparameters and their value is estimated from the reported labels through maximum marginal likelihood estimation. In our spatial setting, reports are sparsely distributed over the area of interest, and consequently each location is unlikely to have multiple observations. Therefore, their approach may suffer from having an arbitrarily large number of free hyperparameters, one for each location, thus making the inference problem computationally infeasible. In contrast, our approach directly models user trustworthiness in the HGP model using a smaller set of parameters, one for each user, and which are easier to estimate from the data using a similar inference approach. In addition, a key difference is that our method can handle reports as continuous estimates rather than single point observations.

The Heteroskedastic GP Model

In this section, we summarise the standard HGP model for spatial regression (see Rasmussen and Williams 2006, for more details). Given a dataset $D = \{(\boldsymbol{x}_{i,j}, y_{i,j})\}$, where $\boldsymbol{x}_{i,j} \in \mathcal{R}^2$ is a two-dimentional location (latitude and longitude) and $y_{i,j} \in \mathcal{R}$ is the value of the i-th observation reported by user j in the location $\boldsymbol{x}_{i,j}$. We want to infer the underlying function $f : \mathcal{R}^2 \rightarrow \mathcal{R}$ which, in our setting, represents the spatial phenomenon observed by the crowd. We assume that $y_{i,j}$ is a noisy sample of f with a zero-mean Gaussian noise $\epsilon_{i,j} \sim \mathcal{N}(0, \sigma)$:

$$y_{i,j} = f(\boldsymbol{x}_{i,j}) + \epsilon_{i,j} \tag{1}$$

where σ is constant across the reporting process. A GP is defined as a distribution over f such that the joint distribution over any subset of function values is multivariate Gaussian. Specifically, the GP distribution over f is specified as:

$$f(\boldsymbol{x}) \sim \mathcal{GP}(m(\boldsymbol{x}), K(\boldsymbol{x}, \boldsymbol{x}'))$$

where $m(\boldsymbol{x}) = \mathbb{E}(f(\boldsymbol{x}))$ is the mean function modelling the expected values of f (often assumed to be constant) and $K(\boldsymbol{x}, \boldsymbol{x}') = cov(f(\boldsymbol{x}), f(\boldsymbol{x}'))$ is the covariance function specifying the correlation between pairs of function values. Both these functions have free hyperparameters controlling the smoothness and the noise properties of the GP estimator. Then, given the conjugate form of the Gaussian likelihood and the GP prior, the inference in GP models yields to a closed form expression for the posterior density over f from which the predictive distribution of the function at different test points can be derived.

While the GP can only model datasets with constant variance noise, HGPs relax this assumption to represent datasets where the noise variances changes across the inputs, i.e. $\epsilon_{i,j} \sim \mathcal{N}(0, \sigma_{i,j})$. This varying noise feature, commonly referred to as *heteroskedasticity*, is particularly relevant to our crowdsourcing settings where data are typically provided by sources with individual noise levels (i.e. the user accuracy). However, unlike the homoskedastic case, heteroskedasticity in GP models makes inference no longer tractable due to the dependency of $\sigma_{i,j}$ on $\boldsymbol{x}_{i,j}$ which does no longer allow a closed form likelihood and leads to an intractable integral for the posterior updates. For this reason, research has focussed on approximate inference in HGP models, using Markov Chain Monte Carlo approaches (Goldberg, Williams, and Bishop 1997) or Expectation-Maximisation (Kersting et al. 2007) and variational Bayes approximation (Lzaro-gredilla and Titsias 2011).

However, a notable tractable exception of HGPs derives from assuming independency between the $\sigma_{i,j}$ terms. That is, the users sample observations with independent noise levels. This assumption is reasonably applicable to the crowdsourcing setting since users typically report observations independently and collusion among crowd members, i.e. groups of users intentionally misreporting their estimates, is not (yet) a primary issue within crowdsourcing systems (Venanzi, Rogers, and Jennings 2013). From this, the likelihood factorises over cases in the dataset and the posterior distribution over the function can be derived as the combination of the HGP kernel and the diagonal noise matrix (see the next section for more detail).

While the standard HGP model can only represents data points with heteroskedastic noise, it does not take into account the different trustworthiness between the users who provide them. Therefore, we now detail our extension to the HGP to model untrustworthy spatial crowd reports.

Crowdsourcing Spatial Functions

In crowdsourcing spatial functions, we collect a number of observations of f submitted by a crowd of n users at different locations. For example, f may represent the environmental process being monitored, such as a weather map, pollution map or radiation map. Thus, the domain of f is the set of locations describing the observed land area and the codomain is the continuous range of values that the function can assume. Each user i provides a set of k_i observations $O_i = \{o_{i,j} : 1 < j < k_i\}$ and let $k = \sum_i^n k_i$ be the total number of observations received from the crowd. Each observation $o_{i,j}$ includes : (i) the user's GPS location $\boldsymbol{x}_{ij} \in \mathcal{R}^2$ (assumed to also be the location of the measurement), (ii) the measured value $y_{i,j} \in \mathcal{R}$ and (iii) the precision $\theta_{i,j} \in \mathcal{R}_+$. In particular, θ is the observed precision modelling the uncertainty in $y_{i,j}$ as reported by the user.

To relate the noise in an observation to the reported precision, we assume that $o_{i,j}$ is a noisy sample of f and the *observed* noise variance (or inverse precision) is specified by $\theta_{i,j}^{-1}$. That is:

$$y_{i,j} = f(\boldsymbol{x}_{i,j}) + \epsilon_{i,j} \qquad \epsilon_{i,j} \sim \mathcal{N}(0, \theta_{i,j}^{-1}) \tag{2}$$

Thus, the equation above specifies the HGP regression problem for multi-user crowd reporting settings. As an example, Figure 1 illustrates a dataset of 30 estimates reported by six users observing the one-dimensional function f (blue-dotted line). Each estimate is plotted as the reported mean y_{ij} (starred points) and the two standard deviations bars $(\pm 2/\sqrt{\theta_{ij}})$ given by the reported precisions.

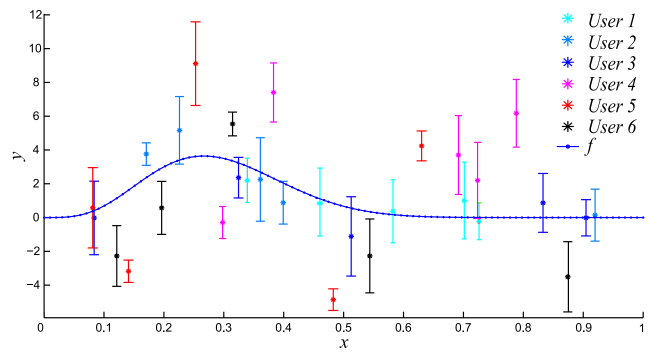


Figure 1: Example dataset with heteroskedastic noise including 30 estimates reported by six users observing the blue-dotted function.

Modelling User Trustworthiness

We characterise the trustworthiness of each user i with an individual parameter $t_i \in [0, 1]$ (1 for a fully trustworthy user and 0 for an untrustworthy user). In particular, we consider user trustworthiness as shaped by the behaviour of a user in reporting inconsistent estimates with respect to f. In practice, our approach models the principle that trustworthy users are expected to sample possibly noisy observations from f. On the other hand, an untrustworthy user can report observations which may be uncorrelated with f and sampled from different statistics. For instance, the previous example in Figure 1 shows that user 6 (black) is potentially untrustworthy since all of its estimates are inconsistent with the true value of f, i.e. $f(x_{i,j}) \notin [y_{i,j} \pm 2/\sqrt{\theta_{ij}}]$. In contrast, user 3 (blue) is more trustworthy since most of its estimates are representative samples of f.

Now, the ability to identify untrustworthy users and handle the inaccuracy of their reports in the data fusion process is required to accurately estimate the function. To address this, we use a trust-based uncertainty scaling technique based on adding extra uncertainty to subsets of data points depending on how much such points are trustworthy. By doing so, the model is able to allow larger variance around untrustworthy points, whilst still modelling correlations in the locality of such points.

More formally, let $\hat{\theta}_{i,j} = t_i \theta_{i,j}$ be the *trusted* precision, i.e. the reported precision linearly scaled by t_i. Then, the regression problem stated in Eq. 1 is updated as follows:

$$y_{i,j} = f(x_{i,j}) + \hat{\epsilon}_{i,j} \qquad \hat{\epsilon}_{i,j} \sim \mathcal{N}(0, (t_i \theta_{i,j})^{-1}) \tag{3}$$

That is, the set of precisions reported by user i is now scaled proportionally to t_i. This produces the effect of increasing the uncertainty in user i's reports up to turning them into completely uninformative contributions when t_i is close to zero.[1] In this way, the model can now refine the data fusion process by filtering untrustworthy estimates depending on the t_i parameters. Thus, the next crucial step is how to learn the values of t_i from the data and how to make predictions of f accordingly.

Trust-Based HGP

To perform inference in the function space, we place a zero-mean GP prior over f, i.e. $m(\boldsymbol{x}) = 0$. Here we use the squared-exponential covariance function which is a commonly used kernel for modelling smoothly varying quantities:

$$K(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f \exp\Big(-\frac{d(\boldsymbol{x}, \boldsymbol{x}')^2}{2l^2}\Big) \quad (4)$$

where d is the line distance between two locations $\boldsymbol{x}$ and $\boldsymbol{x}'$ calculated using the standard equilateral projection:

$$d(\boldsymbol{x}, \boldsymbol{x}') = R_0\sqrt{x^2 + y^2} \quad (5)$$

$$x = (lon - lon')\cos((lat + lat')/2) \quad (6)$$

$$y = lat - lat' \quad (7)$$

where $R_0 = 6,371$ is the mean Earth's radius in kilometers, σ_f is the signal variance and l is the length scale of the squared exponential function.

Recall, in order to have a tractable likelihood, we need to assume independence between the noise terms, i.e. $\hat{\epsilon}_{i,j} \perp \hat{\epsilon}_{i',j'} \Rightarrow \hat{\theta}_{i,j} \perp \hat{\theta}_{i',j'}$ and $t_i \perp t_{i'}$, which is equivalent to assume uncorrelated accuracies between individual measurements and that users are independently trustworthy. Then, to predict the value of f at a new location $\boldsymbol{x}_*$, and let y_* be such a value, let y be the vector of observations, then assuming that y and y_* are Gaussian random vectors, we can write the joint distribution at the test location as:

$$\begin{bmatrix} y \\ y_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(\boldsymbol{x}, \boldsymbol{x}) + \Sigma & K(\boldsymbol{x}, \boldsymbol{x}_*) \\ K(\boldsymbol{x}_*, \boldsymbol{x}) & K(\boldsymbol{x}_*, \boldsymbol{x}_*) \end{bmatrix}\right)$$

where

$$\Sigma = \mathrm{diag}(\hat{\theta}_{i,1}, \ldots, \hat{\theta}_{i,p_i}, \ldots, \hat{\theta}_{n,1}, \ldots, \hat{\theta}_{n,p_n})^{-1}$$

is the $k \times k$ diagonal matrix of the reported precisions, each scaled by the user's t_i parameter. That is, using our trust-based parametrisation of the noise rates, we obtain a joint density with t_i regulating the noise of the user's set of input points.

Next, using the marginalisation properties of the Gaussian distribution, the predictive density of our trust-based HGP (or Trust HGP) is a multivariate Gaussian expressed as follows:

$$p(y_*|\boldsymbol{x}, y, \boldsymbol{x}_*) = \mathcal{N}(E[y_*], \sigma^2(y_*))$$

where

$$E[y_*] = K(\boldsymbol{x}_*, \boldsymbol{x})[K(\boldsymbol{x}, \boldsymbol{x}) + \Sigma]^{-1}y$$

$$\sigma^2(y_*) = K(\boldsymbol{x}_*, \boldsymbol{x}_*) - K(\boldsymbol{x}_*, \boldsymbol{x})[K(\boldsymbol{x}, \boldsymbol{x}) + \Sigma]^{-1}K(\boldsymbol{x}, \boldsymbol{x}_*)$$

are the predictive mean and variance of f at the location $\boldsymbol{x}_*$, respectively, given the hyperparameter set $\Theta = \{\sigma_f, l, t_i, \ldots, t_n\}$.

Then, we can derive the log marginal likelihood by integrating the likelihood over the HGP prior:

$$\mathcal{L} = \ln\Big(\int p(y|f, \boldsymbol{x})p(f|\boldsymbol{x})df\Big) = -\frac{1}{2}y^T C^{-1}y - \frac{1}{2}\ln|C| - \frac{k}{2}\ln(2\pi)$$

where $C = K(\boldsymbol{x}, \boldsymbol{x}) + \Sigma$. The partial derivatives of the likelihood function are:

$$\frac{\partial \mathcal{L}}{\partial \Theta} = \frac{1}{2}y^T C^{-1}\frac{\partial C}{\partial \Theta}C^{-1}y + \frac{1}{2}tr\left(C^{-1}\frac{\partial C}{\partial \Theta}\right)$$

and from Eq. 7, we can find that:

$$\frac{\partial C}{\partial \sigma_f} = 2\sigma_f \exp\Big(-\frac{d^2}{2l^2}\Big) \quad (8)$$

$$\frac{\partial C}{\partial l} = -\frac{\sigma_f^2 d^2}{l^3}\exp\Big(-\frac{d^2}{2l^2}\Big) \quad (9)$$

$$\frac{\partial C}{\partial t_i} = -\frac{1}{t_i^2}\mathrm{diag}(0, \ldots, 0, \theta_{i,1}, \ldots \theta_{i,p_i}, 0, \ldots, 0)^{-1} \quad (10)$$

Given this, we use the maximum marginal likelihood estimator, a standard model selection framework for GP models, to set the values of the hyperparameters, which also include the users' trustworthiness values, i.e. $\Theta_{\mathrm{ML}} = \arg\max_\Theta(\mathcal{L}|\Theta)$. In particular, the analytical gradient of the likelihood with respect to the hyperparameters (Eq. 8, 9, 10) can be used for the efficient search for the maximiser using gradient based optimisation methods.[2]

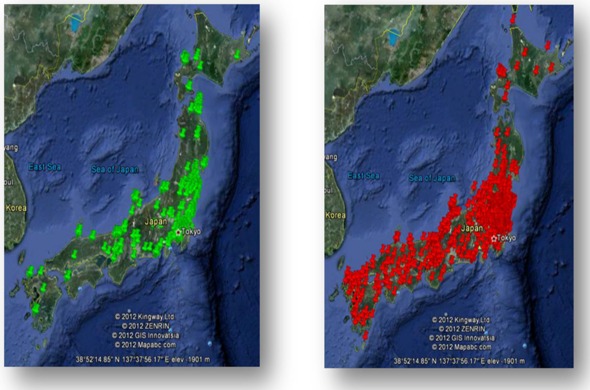
The model training and posterior updates is of time complexity $O(k^3)$. This is the standard complexity of inference in GP methods (Rasmussen and Williams 2006) as a result of the operational cost of inverting the covariance matrix. In practice, we found that our model can handle datasets with up to $2,500$ data points in approximately 5 minutes on a i5 3.6 GHz CPU, 8GB RAM architecture.

Experimental Evaluation

To evaluate our method, we consider the key crowdsourcing application of radiation monitoring where we test the Trust HGP accuracy in making spatial radioactivity predictions against the presence of untrustworthy sensors. Subsequently, we complete our analysis by running simulations on synthetic data which allows us to test the robustness of our method with a number of untrustworthy crowds.

[1]Notice the case $t_i = 0$ produces an infinite value for the variance which is already correctly represented by the IEEE 754 floating point standard and should be handled accordingly in computer programs.

[2]The non-linear conjugate gradient method provided by the `gpml v.2` Matlab toolbox was used in our implementation.



(a) Xively (b) SPEEDI

Figure 2: Image showing the location of the Xively sensors (a) and the SPEEDI sensors (b).

Experimental Setup

In our experiments, we consider the following benchmarks:

- **Standard GP**: The homoskedastic GP (i.e. with a constant-variance noise) with a zero-mean function and a squared exponential covariance function (Rasmussen and Williams 2006, §2.2).
- **HGP**: The standard HGP model without trust parameters, i.e the non-trust version of our model where the trust parameters are statically set to 1, $t_i = 1, \forall i$.
- **Optimal HGP**: This is the *hypothetical* optimal HGP method provided with perfect knowledge of the correct t_i values. That is the Trust HGP, where $t_i = 1$ and $t_i = 0$ for trustworthy and untrustworthy users, respectively, are set in advance. Note we can only make this comparison in the case of the synthetic datasets.

To measure the accuracy of each GP method, we compute the root mean square error (RMSE) with respect to y^*, i.e. the ground truth values of f:

$$RMSE(y, y^*) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - y_i^*)}$$

where N is the total number of predictions. We also consider the *negative continuous rank probability score* (NCRPS) to provide a more comprehensive measure of the probability mass predicted around y_*. This is a non-local scoring rule particularly suitable for scoring predictors providing the properties of *properness* (i.e. the true generative distribution has the best score) and *distance-sensitive* scores (i.e. it is proportional to predictive probability mass placed near the true value) (Kohonen and Suomela 2006). In particular, the NCRPS averaged over N Gaussian predictions is:

	RMSE	NCRPS
Standard GP	30.80 ± 0.30	-64.34 ± 0.04
HGP	64.13 ± 0.99	-9.31 ± 0.12
Trust HGP	26.74 ± 0.27	-7.14 ± 0.08

Table 1: Scores of the predictions of the three GP methods on the Xively dataset.

$$NCRPS(\mathcal{N}(y, \sigma^2), y^*) = \frac{1}{N}\sum_{i=1}^{N}\sigma_i\left(\frac{1}{\sqrt{\pi}} - 2\varphi\left(\frac{y_i^* - y_i}{\sigma_i}\right) - \frac{y_i^* - y_i}{\sigma_i}\left(2\phi\left(\frac{y_i^* - y_i}{\sigma_i}\right) - 1\right)\right)$$

where φ and ϕ denote the probability density function and the cumulative distribution of a standard normal random variable, respectively.

Evaluation on Real-World Data

In this experiment, we present an application of our method to the scenario of crowdsourced radiation monitoring during the Fukushima nuclear disaster. On 3 March 2011, a tsunami caused by a 9 magnitude earthquake hit the east coast of Japan severely damaging the nuclear power plant of Fukushima-Daichii. The subsequent nuclear accident led to radioactivity increases of up to 1,000 times the normal levels in the area of Fukushima and provoked the second-largest world-wide nuclear emergency since Chernobyl, 1985. In response, private individuals deployed 557 Geiger counters across the country (many of them based on open-hardware boars such as Arduino or Goldmine) which were able to report live radiation data through the web connected to the Xively platform (`xively.com`). This entirely crowdsourced Xively sensor network, showed in Figure 3a, came to live in less than two weeks after the disaster and became a key resource for the public to gather live radioactivity information from the disaster scene. However, the key challenge for the rescue teams was to manage the large amount of data streamed by the sensors into a comprehensive spatial radioactivity prediction, considering that an unknown number of unreliable sensors were reporting verifiably wrong measurements. In this scenario, we show how our Trust HGP can be applied to improve the accuracy of radioactivity predictions from the radiation data provided by the Xively network.

We used the readings reported by the Xively sensors over one day, 1 March 2012 (the experiment was repeated over different days with similar results and a live demo of this experiment running on a daily basis is available at `jncm.ecs.soton.ac.uk`). We estimate the mean value y_i and the precision θ_i of each sensor by taking the average and the inverse variance of the series of its measurements. The sensor readings are reported in the unit of microsieverts per hour ($\mu Sv/h$) at an average frequency of 2 readings per hour. In this way, we construct the Xively dataset with 557 reports, one from each sensor, where each report consists of i) x_i the sensor location, ii) y_i, the sensor's average radiation

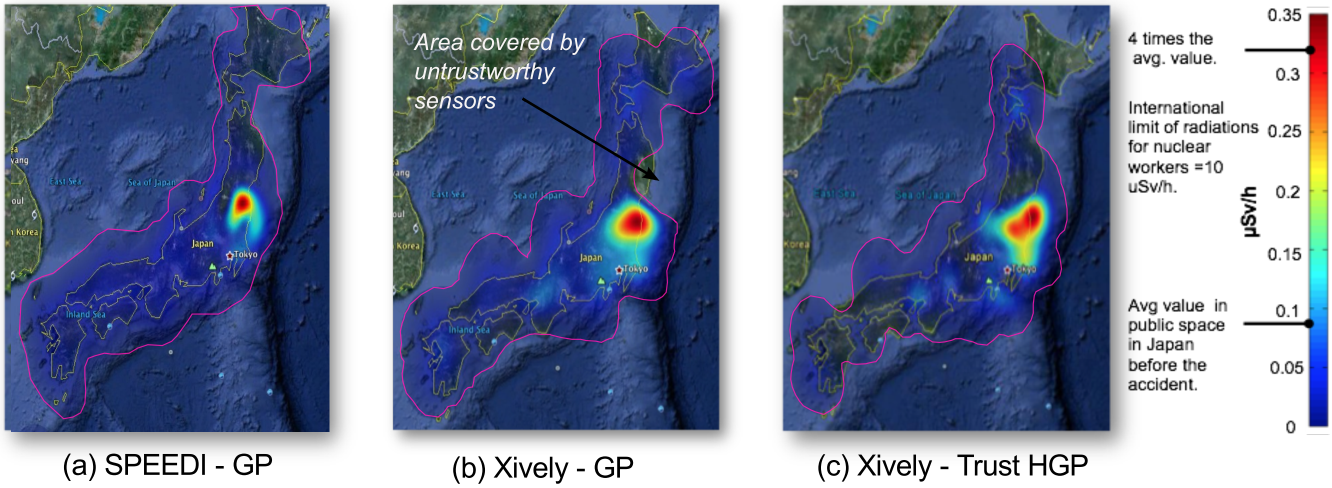


Figure 3: The images show the radiation heat map predicted by the standard GP on the SPEEDI dataset (a) the standard GP on the Xively dataset (b) and the Trust HGP on the Xively dataset (c).

reading, iii)θ_i the sensor's empirical precision. [3]

To build a ground truth for this experiment, we use data provided by the SPEEDI network: the official radiation monitoring network maintained by the Nuclear Division of the Ministry of Science of Japan (MEXT)[4]. The SPEEDI network includes 2122 sensors reporting readings at a frequency of 6 readings per hour, also in the unit of $\mu Sv/h$ (Figure 2b). Thus, we construct a second SPEEDI dataset using the mean and the precision of the readings reported by the SPEEDI sensors. Then, making the reasonable assumption that the SPEEDI dataset are more reliable due to their official source, we run the standard GP on the SPEEDI dataset to generate the ground truth radiation data showed in Figure 3a.

In more detail, Figure 3b and Figure 3c show the predictions of the two methods (GP and Trust HGP) on the Xively dataset depicted as radiation heat maps. While the two predictions are similar in identifying the peak of radioactivity of approximately 0.33 $\mu Sv/h$ near to the location of the Fukishima power plant, they are substantially different in several locations. For example, it can be noticed that the standard GP does not provide valid radiation values near the location of Onagawa in the Miyagi prefecture (38.45 N, 141.44 E). In fact, we manually discovered that some of the sensors located in that area sporadically reported invalid measurements which caused the GP to predict inconsistent radiation values. In contrast, the Trust HGP makes more plausible predictions and overcomes this issue by correctly learning to place a low degree of trustworthiness on such sensors. In particular, it estimated that 17% of the Xively sensors have trustworthiness values lower than 0.5. The same analysis on the SPEEDI sensors revealed that only few of these (less than 1%) were untrustworthy which confirmed our assumption about the SPEEDI network being more reliable.

Finally, Table 1 reports the scores of the predictions of the three methods in $N = 100$ trials. In each run, we randomly sample 80% of the sensors in order to evaluate the performance of the tested methods over different portions of the Xively dataset. The results show that the Trust HGP outperforms the best benchmark by 13% with respect to the RMSE and by 23% with respect to the NCRPS. In more detail, while the HGP improves the NCRPS of the standard GP, the RMSE of the former is significantly worse. In contrast, our method achieves the best performance in both the scores as a result of its correct learning of the trustworthiness values. Thus, this result shows that our method is more accurate and considerably more informative in estimating radiation levels on a prominent crowdsourced spatial dataset.

Evaluation on Synthetic Data

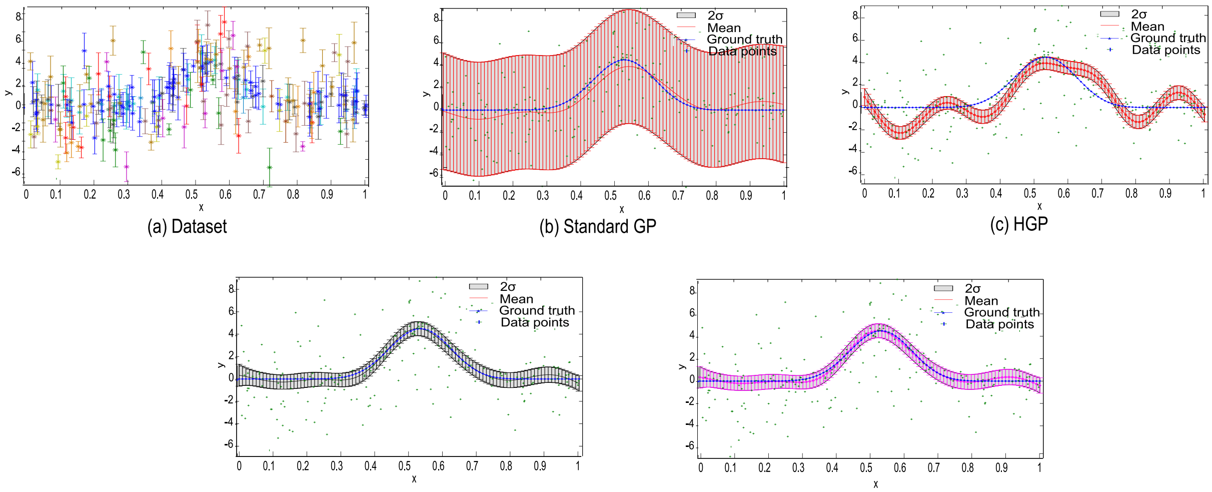
In this experiment, we evaluate the Trust HGP in estimating a one-dimensional function from synthetic reports. We consider a crowd of $n = 20$ users where each user provides k_i observations where $k_i \sim U[3, 20]$. We simulate f using a beta function, $Beta(\alpha, \beta)$ with support in $[0, 1]$ and with random shape parameters sampled as $\{\alpha, \beta\} \sim U[1, 20]$.

To generate synthetic reports of f,we sample the precisions as $\theta_{i,j} \sim \pm U[0.5, 20]$. Then, taking random input point $x_{i,j}$ in the domain of f, the corresponding output $y_{i,j}$ is generated as a Gaussian random sample around the function value $y^*_{i,j}$, i.e. $y_{i,j} \sim \mathcal{N}(y^*_{i,j}, \theta_{i,j})$. Finally, we simulate a percentage ρ of untrustworthy users among the crowd by adding extra noise $w \sim \pm U[1, 5]$ to their set of estimates. In more detail, by randomly sampling between the positive and the negative noise range we avoid the bias of having the noise of untrustworthy estimates always positively or negatively defined.

Figure 4 shows the typical regression of the four methods in this setting. Given a test dataset of 240 estimates with $\rho = 30\%$ (Figure 4a), the standard GP usually produces

[3]This dataset and the Java code to query the Xively sensors are available as supplementary material.

[4]`bousai.ne.jp`



(a) Dataset (b) Standard GP (c) HGP

(d) Trust HGP (e) Optimal HGP

Figure 4: Example of regression with the four GP methods with a dataset of 240 estimates referring to 20 users with $\rho = 30\%$. Specifically, the ground truth f is the blue-dotted line and the GP predictions are depicted as the mean (red line) and the 2σ shaded area.

good mean-value predictions but overestimates the uncertainty (Figure 4b). This also agrees with the empirical findings by Kersting et al. (2007) from a general evaluation of GPs applied to heteroskedastic settings. Instead, the HGP predictions have lower uncertainty but are less accurate as the mean-value prediction is typically far from f (Figure 4c). Furthermore, the irregular shape of the HGP's predictive function is explained by the effect of chasing every noisy point due to considering all the reports as equally trustworthy. In contrast, the Trust HGP achieves the best trade-off between high accuracy and low predictive uncertainty (Figure 4d) and its regression is almost identical to the one of the optimal HGP (Figure 4e). In fact, the correct learning of the trustworthiness values enables our method to exclude most of the untrustworthy points by placing a high noise around these points.

In more detail, Figure 5 shows the performance of the four methods in $N = 200$ repeated runs varying ρ from 0% to 60%. The graph shows that the Trust HGP outperforms the best benchmark by up to 34% in the RMSE (Figure 5a) and up to 70% in the NCRPS (Figure 5b). In particular, it performs close to the optimal HGP up to $\rho = 30\%$ and, after this point, it's accuracy gradually conforms to the other methods as ρ increases. This means that the Trust HGP can correctly handle crowds with a moderately large presence of untrustworthy users. Specifically, the error in its predictions is only 25% worse than the Optimal HGP for $\rho = 50\%$, and it is almost zero when the majority of trustworthy users within the crowd is more than 70%.

Furthermore, the NCRPS shows that Trust HGP's predictions are significantly more accurate and with low uncertainty, hence very informative. Also, of note is the fact that the HGP outperforms the standard GP in terms of NCRPS in any ρ configuration, while the latter typically has a lower RMSE. However, both of these methods are less accurate than the Trust HGP.

Conclusions

In this paper, we addressed the problem of learning continuous functions from crowdsourced spatial data using a trust-based HGP modelling approach. The key innovation of our approach lies in combining an HGP with a user trust model introducing a set of trust hyperparameters to model the different accuracies of the users in reporting their estimates. In particular, by training our model with the reports gathered from the crowd, we are able to estimate the underlying spatial function at new locations and also learn the trustworthiness level of each user. Furthermore, we showed that our methods significantly improves, the quality of the predictions of the standard GP and HGP methods by up to 23% in the key disaster response application of crowdsourced radiation monitoring using real-world data from the 2011 Fukushima nuclear disaster. We also evaluate our method on synthetic data showing that it outperforms the benchmarks by up to 70% and is robust against an up to 30% presence of untrustworthy users. Therefore, our method is able to provide an informative support to decision makers to act upon crowdsourced information.

These results open several directions for future work. First, we would like to explore settings in which user trustworthiness levels are no longer independent which may lead to coalitions of crowd members with similar behaviours. Second, we would like to incorporate temporal dynamics into our model which will make it potentially more interesting for a broader class of space-time dependent crowdsourcing settings.

Acknowledgments

The authors gratefully acknowledge funding from the UK Research Council for the ORCHID project, grant EP/I011587/1.

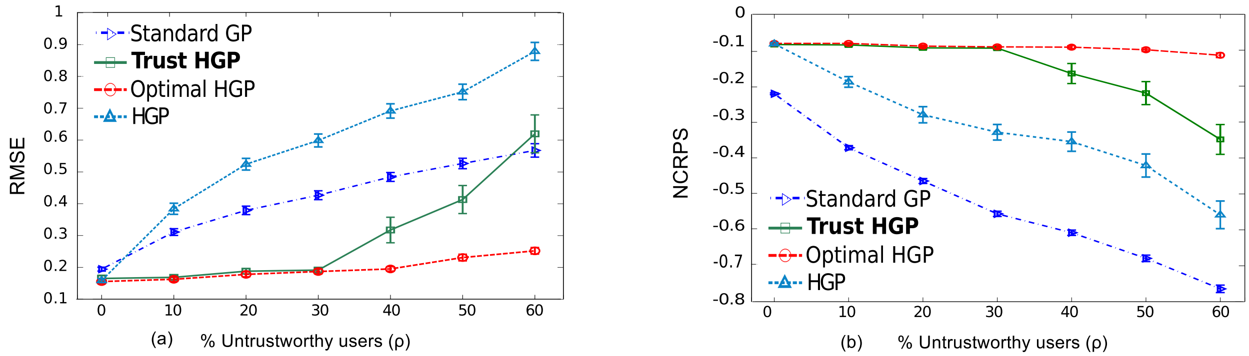


Figure 5: Performance of the four methods measured by the RMSE (a) and the NCRPS (b).

References

Bachrach, Y.; Graepel, T.; Kasneci, G.; Kosinski, M.; and Van Gael, J. 2012. Crowd iq-aggregating opinions to boost performance. *choice* 9:8.

Dawid, A., and Skene, A. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics* 20–28.

Faulkner, M.; Olson, M.; Chandy, R.; Krause, J.; Chandy, K. M.; and Krause, A. 2011. The next big one: Detecting earthquakes and other rare events from community-based sensors. In *Information Processing in Sensor Networks (IPSN), 2011 10th International Conference on*, 13–24. IEEE.

Gertz, E., and Di Justo, P. 2012. *Environmental Monitoring with Arduino: Building Simple Devices to Collect Data About the World Around Us*. O'Reilly Media, Inc.

Goldberg, P.; Williams, C.; and Bishop, C. 1997. Regression with input-dependent noise: A gaussian process treatment. *Advances in neural information processing systems* 10:493–499.

Groot, P.; Birlutiu, A.; and Heskes, T. 2011. Learning from multiple annotators with gaussian processes. *Artificial Neural Networks and Machine Learning–ICANN 2011* 159–164.

Hall, D., and Jordan, J. 2010. *Human-centered information fusion*. Artech House Publishers.

Herring, R.; Hofleitner, A.; Amin, S.; Nasr, T.; Khalek, A.; Abbeel, P.; and Bayen, A. 2010. Using mobile phones to forecast arterial traffic through statistical learning. In *89th Transportation Research Board Annual Meeting, Washington DC*.

Horvitz, E. J.; Apacible, J.; Sarin, R.; and Liao, L. 2012. Prediction, expectation, and surprise: Methods, designs, and study of a deployed traffic forecasting service. *arXiv preprint arXiv:1207.1352*.

Kamar, E.; Hacker, S.; and Horvitz, E. 2012. Combining human and machine intelligence in large-scale crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, 467–474. International Foundation for Autonomous Agents and Multiagent Systems.

Kersting, K.; Plagemann, C.; Pfaff, P.; and Burgard, W. 2007. Most likely heteroscedastic gaussian process regression. In *Proceedings of the 24th international conference on Machine learning*, 393–400. ACM.

Kohonen, J., and Suomela, J. 2006. Lessons learned in the challenge: making predictions and scoring them. *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment* 95–116.

Krause, A.; Horvitz, E.; Kansal, A.; and Zhao, F. 2008. Toward community sensing. In *Proceedings of the 7th international conference on Information processing in sensor networks*, 481–492. IEEE Computer Society.

Lzaro-gredilla, M., and Titsias, M. K. 2011. Variational heteroscedastic gaussian process regression. In *In 28th International Conference on Machine Learning (ICML-11*, 841–848. ACM.

Rasmussen, C., and Williams, C. 2006. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge, MA.

Raykar, V. C.; Yu, S.; Zhao, L. H.; Valadez, G. H.; Florin, C.; Bogoni, L.; and Moy, L. 2010. Learning from crowds. *The Journal of Machine Learning Research* 99:1297–1322.

Reece, S.; Roberts, S.; Claxton, C.; and Nicholson, D. 2009. Multi-sensor fault recovery in the presence of known and unknown fault types. In *Information Fusion, 2009. FUSION'09. 12th International Conference on*, 1695–1703. IEEE.

Sadilek, A.; Kautz, H. A.; and Silenzio, V. 2012. Modeling spread of disease from social interactions. In *ICWSM*.

Venanzi, M.; Rogers, A.; and Jennings, N. R. 2013. Trust-based fusion of untrustworthy information in crowdsourcing applications. In *12th Int. Conference on Autonomous Agents and Multi-Agent Systems, AAMAS 2013*.