Estimating Predictive Variance for Statistical Gas Distribution Modelling

Achim J. Lilienthal, Sahar Asadi and Matteo Reggente

AASS Research Center, Örebro University, Sweden

Abstract. Recent publications in statistical gas distribution modelling have proposed algorithms that model mean and variance of a distribution. This paper argues that estimating the predictive concentration variance entails not only a gradual improvement but is rather a significant step to advance the field. This is, first, since the models much better fit the particular structure of gas distributions, which exhibit strong fluctuations with considerable spatial variations as a result of the intermittent character of gas dispersal. Second, because estimating the predictive variance allows to evaluate the model quality in terms of the data likelihood. This offers a solution to the problem of ground truth evaluation, which has always been a critical issue for gas distribution modelling. It also enables solid comparisons of different modelling approaches, and provides the means to learn meta parameters of the model, to determine when the model should be updated or re-initialised, or to suggest new measurement locations based on the current model. We also point out directions of related ongoing or potential future research work.

Keywords: Gas distribution modelling; gas sensing; mobile robot olfaction; density estimation, model evaluation.
PACS: 01.30.Cc

1. INTRODUCTION

Gas distribution modelling (GDM) is the task of deriving a truthful representation of the observed gas distribution from a set of spatially and temporally distributed measurements. It is very challenging mainly since in many realistic scenarios gas is dispersed by turbulent advection, which creates a concentration field of fluctuating, intermittent patches of high concentration [1].

We can distinguish two types of GDM approaches: model-based and model-free. Model-based approaches infer the parameters of an analytical gas distribution model from the measurements. In principle, Computational Fluid Dynamics (CFD) models can be applied, which solve the governing equations numerically. They are, however, computationally very expensive, become intractable for higher resolutions in typical real world settings and depend sensitively on accurate knowledge of the environment state (boundary conditions), which is not available in practical situations. Many model-based approaches were developed for atmospheric dispersion [2]. Such models typically cannot efficiently incorporate sensor information on the fly and do not provide a sufficient level of detail. This is important since critical gas concentrations often have a local character in complex settings. Simpler analytical models such as [3] often rest on rather restrictive assumptions.

In this paper, we consider a class of model-free approaches, which create a statistical model of the observed gas distribution. In a pure form, these *statistical approaches to distribution modelling* treat input data as random variables and derive a statistical description without making strong assumptions on the functional form of the distribution. This includes that they do not assume certain environmental conditions (such as a uniform airflow, for example). Statistical approaches offer complementary strengths compared to model-based approaches: they do not rely on the validity of the underlying physical model, can provide a higher resolution, are computationally less expensive and generally less demanding in terms of the required knowledge about the state of the environment. Most of the available algorithms, discussed in Sec. 2, create a two dimensional spatial model that represents time-constant structures in the gas distribution in terms of the distribution mean. Three recently proposed approaches, introduced in Sec. 2.1, model not only the mean but also the variance of the distribution. In the following, we argue that this entails a significant improvement for statistical gas distribution modelling.

2. STATISTICAL GDM

This section reviews statistical GDM methods developed for mobile robots and tested at small scales.

A common approach to creating a representation of a time-averaged concentration field is to acquire measurements using a fixed grid of gas sensors over a prolonged period of time and to map average [3] or peak [4] concentrations obtained to the given grid approximation of the environment. Consecutive measurements with a single sensor were used in [5]. To make predictions at locations different from the measurement points, bi-cubic interpolation was applied in the case of equidistant measurements and triangle-based cubic interpolation in the

case of non-equidistant measurements. A problem with such interpolation methods is that there is no means of "averaging out" instantaneous response fluctuations. Response values that were measured very close to each other appear independently in the gas distribution map and thus the representation tends to get more and more jagged while new measurements are added.

Histogram methods reflect the spatial correlation of concentration measurements to some degree by the quantization into histogram bins. The 2-d histogram proposed in [6] accumulates the number of "odor hits" received in an area assigned to the histogram bins. Odor hits are counted whenever the response of a gas sensor exceeds a defined threshold. Disadvantages of this method include the dependency on bin size and selected threshold, that a perfectly even coverage of the inspected area is required, and that only binary information is used and so useful information is discarded.

Kernel extrapolation distribution mapping (Kernel DM) is inspired by non-parametric estimation of density functions using a Parzen window and can be seen as an extension of histogram methods. The concentration field is represented in the form of a grid map. Spatial integration is carried out by convolving sensor readings and modelling the information content of the point measurements with a Gaussian kernel [7].

2.1. Gaussian Process Mixture GDM

None of the methods discussed so far models concentration fluctuations. The enhanced Kernel DM+V algorithm [8], detailed in Sec. 2.2, also estimates the observed distribution variance. Another method that predicts mean and concentration variance uses Gaussian process mixture (GPM) models [9]. It treats GDM as a regression problem. Two components of the GPM represent background signal and areas of high concentration. The components of the mixture model and a gating function, that decides to which component a data point belongs, are learned using Expectation Maximization. In contrast to Kernel DM+V, the model is represented directly using the training data. Because it requires the inversion of matrices that grow with the number of training samples n, the computational complexity of learning the GPM is $\mathcal{O}(n^3)$. This is addressed in [9] by adaptive sub-sampling of the observations to obtain a sparse training set. Similarly to Kernel DM+V, the dependency between nearby locations is modelled in the GPM approach by a radially symmetric, squared exponential covariance function.

2.2. Kernel DM+V

For the illustrating examples in Sec. 3 we use Kernel DM+V [8] to compute distribution models. The performance of this algorithm was found to be level with the GPM approach [9, 8] (see Sec. 3.2) but has the advantages of simplicity, lower computational complexity, and that it is more generally applicable.

Kernel DM+V uses a uni-variate Gaussian weighting function $\mathcal{N}$ to represent the importance of measurement r_i obtained at location $\mathbf{x}_i$ with respect to the measurement statistics over time at grid cell k. First, two temporary grid maps are computed – $\Omega^{(k)}$ by integrating importance weights and $R^{(k)}$ by integrating weighted readings:

$$\begin{aligned} \Omega^{(k)} &= \textstyle\sum_{i=1}^{n} \mathcal{N}(|\mathbf{x}_i - \mathbf{x}^{(k)}|, \sigma), \\ R^{(k)} &= \textstyle\sum_{i=1}^{n} \mathcal{N}(|\mathbf{x}_i - \mathbf{x}^{(k)}|, \sigma) \cdot r_i. \end{aligned} \tag{1}$$

Here, $\mathbf{x}^{(k)}$ denotes the center of cell k and the kernel width σ is a parameter of the algorithm. The integrated weights $\Omega^{(k)}$ are used for normalisation of the weighted readings $R^{(k)}$ (thus even coverage is not necessary) and to compute a further map $\alpha^{(k)}$, which estimates the confidence in the obtained estimates. The confidence map is used to compute the mean concentration estimate $r^{(k)}$ as

$$\begin{aligned} \alpha^{(k)} &= 1 - e^{-(\Omega^{(k)})^2/\sigma_\Omega^2} \\ r^{(k)} &= \alpha^{(k)} \tfrac{R^{(k)}}{\Omega^{(k)}} + \{1 - \alpha^{(k)}\} r_0 \end{aligned} \tag{2}$$

where r_0 represents the mean concentration estimate for cells where we do not have sufficient information from nearby readings, indicated by a low value of $\alpha^{(k)}$. Currently, we set r_0 to be the average over all sensor readings. The scaling parameter σ_Ω^2 defines a soft margin for values of $\Omega^{(k)}$. Similarly to the distribution mean map, Eq. (2), the variance map $v^{(k)}$ is computed from *variance contributions* integrated in a further temporary map $V^{(k)}$

$$\begin{aligned} V^{(k)} &= \textstyle\sum_{i=1}^{n} \mathcal{N}(|\mathbf{x}_i - \mathbf{x}^{(k)}|, \sigma)(r_i - r^{(k(i))})^2, \\ v^{(k)} &= \alpha^{(k)} \tfrac{V^{(k)}}{\Omega^{(k)}} + \{1 - \alpha^{(k)}\} v_0 \end{aligned} \tag{3}$$

where $k(i)$ is the cell closest to the measurement point $\mathbf{x}_i$, and v_0 is an estimate of the distribution variance in regions far from measurement points, computed here as the average over all variance contributions.

3. GDM EVALUATION

Ground truth evaluation has always been a critical issue for gas distribution modelling with mobile robots. The capability to identify hidden parameters, for example the location of the gas source, has been used to test gas distribution models. However, the distance of the distribution maximum to the gas source can only serve as a rough approach to validate the distribution model. Considering only fixed measurement points, a feasible experimental set-up would be to use a stationary grid of gas sensors and to compare the model derived from all but

one or a few sensors with the measurements of the left out sensors. We apply a similar method here and create the model using a sub-set of measurements obtained with a mobile robot and compare the model predictions with unseen measurements also obtained with the robot.

In our experiments the robot followed a sweeping trajectory. It was driven at a maximum speed of 5 cm/s and periodically stopped at pre-defined points, constantly acquiring measurements at a rate of 0.8 Hz. The gas source was a small cup filled with ethanol. Apart from a SICK laser range scanner for pose correction, the robot was equipped with a Sensirion SHT11 digital humidity/temperature sensor and six Figaro gas sensors enclosed in an aluminum tube, actively ventilated through a fan. The tube was horizontally mounted at the front side of the robot at a height of 34 cm (see Fig. 1). We consider only the output of one TGS 2620 sensor here.

An obvious way to measure how well unseen measurements are predicted by the distribution model is to compute the average prediction error. Due to the large fluctuations of the instantaneous gas distribution, however, this measure of model quality is not particularly suitable for gas distribution modelling. A gas distribution model should represent the time-averaged concentration *and* the expected fluctuations. These properties are both captured by the negative log predictive density (NLPD), which is a standard criterion to evaluate distribution models. Under the assumption of a Gaussian posterior $p(r_i|\mathbf{x}_i)$, the NLPD of unseen measurements $\mathscr{D} = \{r_1, ..., r_n\}$ acquired at locations $\{x_1, ..., x_n\}$ is computed as

$$NLPD = -\frac{1}{n}\sum_{i\in\mathscr{D}} \log\{p(r_i|\mathbf{x}_i)\} = \frac{1}{2n}\sum_{i\in\mathscr{D}} \left\{ \log \hat{v}(x_i) + \frac{(r_i - \hat{r}(x_i))^2}{\hat{v}(x_i)} \right\} + \frac{1}{2}\log(2\pi). \quad (4)$$

An estimate of the predictive variance is required to compute the NLPD. The importance of including the predictive variance into the criterion for the quality of gas distribution models can be seen in Fig. 1, which shows a comparison of a gas and a temperature model created from measurements recorded with a mobile robot along a sweeping path. The plots in the left part of the figure were created by computing the distribution model up to a certain time and comparing it to the true unseen values (red circles). Model predictions are indicated as predictive mean $\pm 3\times$predicted standard deviation. This comparison demonstrates that gas distribution models typically exhibit more pronounced spatial variance variations while the information is mainly located in the predictive mean in case of the temperature distribution model.

3.1. Learning meta parameters

As an example of an algorithm that provides an estimate of the predictive variance we use Kernel DM+V. It depends mainly on the meta-parameters kernel width σ and cell size c. Based on the NLPD defined in Eq. (4), we learn these meta parameters by dividing the samples $\mathscr{D}$ into disjoint sets $\mathscr{D}_{train}$ and $\mathscr{D}_{test}$ and determine optimal values of the model parameters by cross-validation on $\mathscr{D}_{train}$, keeping $\mathscr{D}_{test}$ for evaluation. Since we use Kernel DM+V we have $\hat{v}(x_i) = v^{(k(i))}$, $\hat{r}(x_i) = r^{(k(i))}$ in Eq. (4).

3.2. Comparison of GDM Approaches

In the same way, in which we evaluate a fixed distribution model depending on its meta parameters, we can compare different GDM approaches by comparing the respective NLPD for unseen measurements. Since the goal is to maximize the likelihood of unseen data, we will prefer models that minimise the NLPD. Tab. 1 shows a NLPD comparison of the GPM method (see Sec. 2.1) with Kernel DM+V (see Sec. 2.2) based on data sets from three different environments in which the robot carried out a sweeping movement consisting of two full sweeps (for details see [9]). The first sweep was used for training and the second sweep (in opposite direction) for testing. As a preliminary result from this investigation we find that GPM and Kernel DM+V offer a comparable performance for gas distribution modelling in the considered environments. This gives Kernel DM+V a slight edge because it scales better to larger numbers n of data samples (having complexity $\mathscr{O}[n \cdot (\frac{\sigma}{c})^2]$ compared to $\mathscr{O}[n^3]$) and the fact that the learning procedure is simpler.

Efficient approaches to GDM will probably apply some form of sub-sampling in a first stage. Again, the predictive variance via the NLPD allows for a meaningful comparison of different sub-sampling strategies.

3.3. Time-dependent GDM

A crucial assumption that we make when building stationary statistical models is that the statistical description is learned from measurements that are generated by a time-constant random process. It is clear that this assumption is not generally valid. We can address this issue in several ways. Regression approaches such as GPM could be extended by the dimension of time. The Kernel DM+V algorithm that is based on the idea of density estimation could be extended by recency weights or a method that detects when and how to change the time-window over which the distribution model is computed. A measure such as the NLPD does not only allow to compare different approaches to time-dependent GDM,

TABLE 1. Comparison of Kernel DM+V and GPM.

Dataset	NLPD, GPM	NPLD, Kernel DM+V
3-rooms	-1.54	-1.44
corridor	-1.60	-1.81
outdoor	-1.77	-1.75

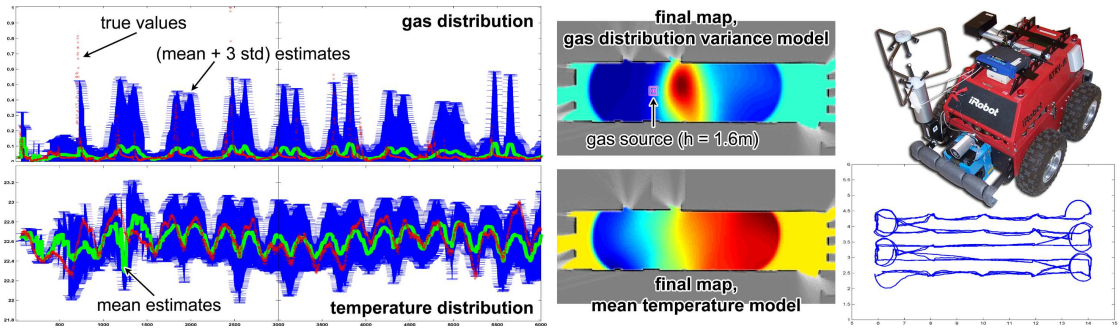

FIGURE 1. Comparison between gas and temperature model recorded with a robot along a sweeping path (shown on the right side). The models are visualised in terms of concatenated predictions of the next 60s of unseen measurements based on the model computed from all the measurements before (best viewed in color).

it also provides a means to detect when the time-window has to be modified and to decide how. An efficient solution might be a "lazy update" mechanism. The quality of the current distribution map could be continuously evaluated on new sensor readings and updated only if the data likelihood drops significantly. An update could also be compared to a model computed from fewer, more recent samples only. By eventually selecting this map, the representation could follow slow distribution changes over time. However, the main benefit of such an approach would be in efficiency of the algorithm since it would still not be possible to extrapolate on time-dependent trends.

3.4. Sensor Planning

A statistical distribution model can be considered good or *truthful* if it explains the measurements, which were used to build the model, and predicts new observations well. To obtain a truthful representation, we need to consider a *sufficient* number of measurements. To quantify this requirement, we can again use the NLPD to compute internal consistency (how well are training data explained) and predictive power of the model (how well are unseen measurements predicted).

A related question is at which locations the next measurements should be carried out in order to obtain a good model in minimum time. Again, the NLPD can be used to compare different sensing strategies regarding their suitability for gas distribution modelling. Further, the predictive variance is an important ingredient for techniques that suggest new measurement locations based on the current model (sensor planning). Appropriate sensor planning strategies need to be evaluated and it is to be expected that they will employ a cost function, which prioritises measurements in areas with high uncertainty, high concentration or high variance.

4. CONCLUSIONS

This paper argues that recently proposed gas distribution modelling algorithms (see Sections 2.1 and 2.2) entail more than a gradual improvement by providing an estimate of the predictive concentration variance. First, estimating the predictive variance captures the particular structure of gas distributions, which exhibit strong fluctuations with considerable spatial variations as a result of the intermittent character of gas dispersal. Accordingly, it is important to consider this feature of gas distributions for model evaluation. A second substantial step forwards is thus that the predictive variance allows to compute the negative log predictive density (NLPD) as a more meaningful measure to evaluate distribution models. The NLPD offers a solution to the problem of ground truth evaluation, which has always been a critical issue for gas distribution modelling. It not only enables solid comparisons of different modelling and sensor planning approaches, but also provides the means to learn meta parameters of the model, to determine when the model should be updated or re-initialised, or to suggest new measurement locations based on the current model.

REFERENCES

1. B. Shraiman, and E. Siggia, *Nature* **405**, 639–646 (2000).
2. H. Rørdam et al., Regulatory odour model development: Survey of modelling tools and datasets with focus on building effects, Tech. Rep. 541 (2005).
3. H. Ishida, T. Nakamoto, and T. Moriizumi, *Sensors and Actuators B* **49**, 52–57 (1998).
4. A. Purnamadjaja, and R. Russell, "Congregation Behaviour in a Robot Swarm Using Pheromone Communication," in *Proc. of the Australian Conf. on Robotics and Automation*, 2005.
5. P. Pyk et al., *Auton Robot* **20**, 197–213 (2006).
6. A. Hayes, A. Martinoli, and R. Goodman, *IEEE Sensors Journal, Special Issue on Electronic Nose Technologies* **2**, 260–273 (2002).
7. A. J. Lilienthal, and T. Duckett, *Robotics and Autonomous Systems* **48**, 3–16 (2004).
8. A. Lilienthal et al., "A Statistical Approach to Gas Distribution Modelling with Mobile Robots, The Kernel DM+V Algorithm," 2009, submitted to IROS.
9. C. Stachniss, C. Plagemann, and A. Lilienthal (2009), to appear.