We thank all reviewers for their useful comments and positive feedback. We’ll fix all minor comments and typos.

Reviewer (R) 1 *“The paragraph in the introduction reads very technical (...)”* In the Introduction, we focus on GP models for aggregated data and multiple instance learning whereas the Related work section mentions work beyond GPs. At the time of submission, we didn’t know of any other multi-task GP model for aggregated data. There are two recent submissions to arxiv (see reply to R2) that we’ll use to follow the reviewer’s recommendation. *“Section 2 was straightforward to follow up to 2.3 (SVI). (...)”* Due to space restrictions we summarised the theory from Section 2.3 since this piece of literature has been described before in Moreno-Muñoz et al. (2018). We’ll expand the description of Section 2.3 in the Appendix. *“l.145ff: The notation has confused me when first reading, (...)”* In L132 we use $\mathbf{y}(\upsilon)$ to refer to the vector of outputs as a function of υ and in L145 $\mathbf{y}$, without the argument υ, to refer to the output vector of the dataset. We’ll clarify in the paper. *“Also l.155: do you mean the process f (...)”* Yes, we refer to f *“l.150ff: How are the inducing inputs Z chosen? (...)”* We use k-means over the input data with $k = M$, the number of inducing inputs. We fix them during optimisation and assume the inducing inputs are points, but we could have also defined them as intervals or supports. *“(...) computational complexity?”* Similar to the one in Moreno-Muñoz et al. (2018): $\mathcal{O}(QM^3 + JNQM^2)$, where J depends on the type of likelihood. *“synthetic data: Could you give an example (...)?”* These could represent two histograms, for example, defined over bins with different sizes. *“what is meant by "support data" (...)”* One-unit support data: data with a support of one unit. *“predicted training count data"”* Predictions made by the trained model over the training data. *“what you mean by 5x5”* a squared support of 5 years for the input `age` times 5 years for the input `years` of the study. *“Now that the likelihood is Gaussian, why not go for exact inference”* That’s true, but this wouldn’t work in the general case, for example, we couldn’t apply this for the toy example. *“Figure 3: I don’t understand the red line:”* It is the ground truth obtained directly by the sensor. We’ll clarify this in the new version. *“Do you have a thought (...) sensors (...)”* An idea previously used in other papers is to assume that each spatial location is a different output. We’re looking into this for our application in air pollution. *“Extend explanation in Section 2.3 (...)” “Rewrite the section on experiments (...). ”* We’ll do this as explained above. **Reviewer 2** *“graphical models(..).”* We’ll add graphical models to the final version. *“bars in Figure 1 (...)”* These are meant to be read as histograms. We’ll add another plot zooming in the prediction range. *“(...) assumption that ‘the correlation between tasks will remain constant (...)’ ”* Our most sincere apologies. This is in no way straightforward and will involve a model along the lines of the Gaussian process regression networks (Wilson et al, 2011). We will remove this statement from the manuscript. *“(...) availability of the code”* We have our code on GPy, and we’ll make it available after the decision. *“P.S. A paper having similar goals appeared on arXiv (...)”* Two papers with similar goals appeared on arXiv recently, the one mentioned by the reviewer, “Multi-resolution Multi-task GPs” (arxiv1) and “Spatially Aggregated GPs with Multivariate Areal Outputs” (arxiv2). Differences: we allow heterogeneous likelihoods (compared to arxiv1 and arxiv2), an exact solution to the integration of the latent function through the kernel in Smith et al (2018) (different to arxiv1); and inducing inputs for computational complexity (different to arxiv2). We’ll add these references. *“(...) discussion regarding multi-fidelity methods (...)”* Very relevant, thanks. We’ll add this to the discussion. *“If possible, adding a more involved experiment (...)”* See reply to R3 (experiment on more tasks). **Reviewer 3** *“ (...) mention several related multi-task GPs (e.g., [1],[2],[3]).”* We’ll add the references. [1] and [3] are particular cases of LMC as it has been described in detail by Alvarez et al (2012). *“(...) related work missing here: [4] (...) differences and advantages (...)”* [4] does not attempt to do simultaneous prediction of several variables, only one variable is considered. They mainly use GPs for creating data from different auxiliary sources. Other differences: they only consider Gaussian regression and they do not include inducing variables. *“the data might be aggregated by another procedure, e.g., simple summation or population weighted average; ”* Agreed. Our motivation was to have a general purpose model. Other types of aggregation will require prior knowledge by the user. This can be extended in future work. *“(...) aggregation (...) at the likelihood level?”* It’s happening in a sense because the latent functions obtained after the integration modulate the parameters of each likelihood in different ways, depending on a_{d_q}. *“(...) I think it would be more efficient to estimate $a_{d,q}$ instead of B_q.”* We estimate $\mathbf{B}_q$ by estimating first the Cholesky factor $\mathbf{L}_q$. See L187. This is efficient. *“the proposed model should be compared with any typical baseline”* See Fig 1. We’ll include the 5×5 resolution case and the SNLP metric in the new version and similar baselines for the other experiments (toy and air pollution). *“experimental results considering more tasks”* We ran another experiment with the Fertility dataset: four outputs (two high-res few data points, two low-res many more data points) and compared two versions of our model: all outputs as Gaussians and all outputs as heteroscedastic Gaussians. SMSEs for both models are comparable, but the model with heteroscedastic Gaussians outperforms in terms of the SNLP. We’ll add this experiment to the new version with many more details. *“*resolution 5×5* (...)”* See reply to R1.

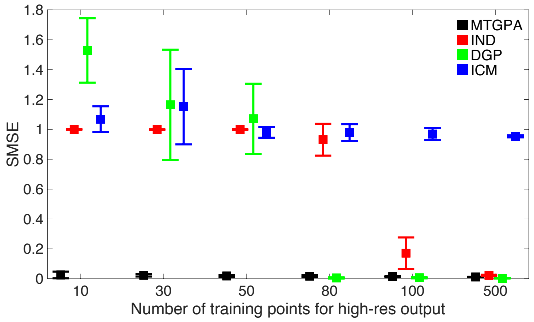

Figure 1: Fertility rates 2×2 low resolution case. MTGPA (our method); IND (Independent GP with aggregated inputs); DGP (Dependent GPs, ref [2], R3); ICM (Intrinsic Co-regionalisation Model or Multi-task GPs, ref [3], R3). DGP and ICM use the centroid of the area as input. MTGPA performs better or similar to baselines as we increase the number of training points for the high-res output.