Visualization in Bayesian workflow

Jonah Gabry†
Department of Statistics and ISERP, Columbia University, New York, USA.
E-mail: jgabry@gmail.com
Daniel Simpson†
Department of Statistical Sciences, University of Toronto, Canada.
Aki Vehtari
Department of Computer Science, Aalto University, Espoo, Finland.
Michael Betancourt
ISERP, Columbia University, and Symplectomorphic, LLC, New York, USA.
Andrew Gelman
Departments of Statistics and Political Science, Columbia University, New York, USA.

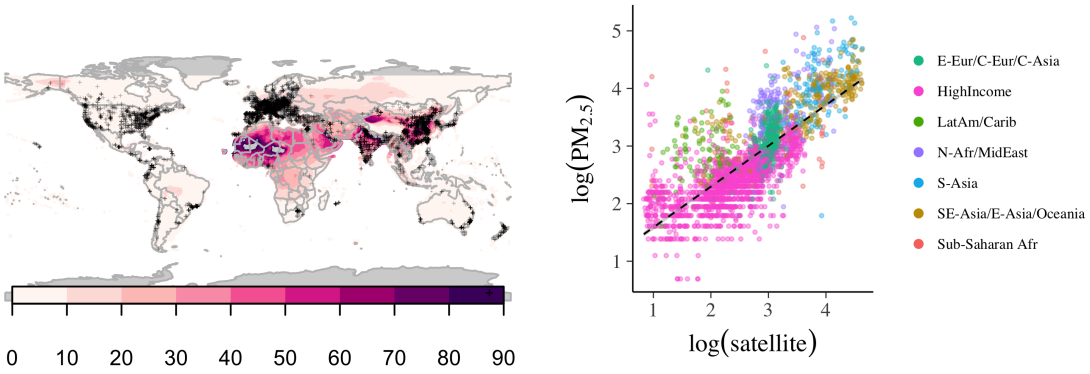
Summary. Bayesian data analysis is about more than just computing a posterior distribution, and Bayesian visualization is about more than trace plots of Markov chains. Practical Bayesian data analysis, like all data analysis, is an iterative process of model building, inference, model checking and evaluation, and model expansion. Visualization is helpful in each of these stages of the Bayesian workflow and it is indispensable when drawing inferences from the types of modern, high-dimensional models that are used by applied researchers.

1. Introduction and running example

Visualization is a vital tool for data analysis, and its role is well established in both the exploratory and final presentation stages of a statistical workflow. In this paper, we argue that the same visualization tools should be used at all points during an analysis. We illustrate this thesis by following a single real example, estimating the global concentration of a certain type of air pollution, through all of the phases of statistical workflow: (a) Exploratory data analysis to aid in setting up an initial model; (b) Computational model checks using fake-data simulation and the prior predictive distribution; (c) Computational checks to ensure the inference algorithm works reliably, (d) Posterior predictive checks and other juxtapositions of data and predictions under the fitted model; (e) Model comparison via tools such as cross-validation.

The tools developed in this paper are implemented in the `bayesplot` R package (Gabry, 2017; R Core Team, 2017), which uses `ggplot2` (Wickham, 2009) and is linked to—though not dependent on—Stan (Stan Development Team, 2017a,b), the general-purpose Hamiltonian Monte Carlo engine for Bayesian model fitting.

†Joint first author



(a) The satellite estimates of $PM_{2.5}$. The black points indicate the locations the ground monitors.

(b) A scatterplot of $\log(PM_{2.5})$ vs log(satellite). The points are colored by WHO super region.

Fig. 1: *Data displays for our running example of exposure to particulate matter.*

In order to better discuss the ways visualization can aid a statistical workflow we consider a particular problem, the estimation of human exposure to air pollution from particulate matter measuring less than 2.5 microns in diameter ($PM_{2.5}$). Exposure to $PM_{2.5}$ is linked to a number of poor health outcomes, and a recent report estimated that $PM_{2.5}$ is responsible for three million deaths worldwide each year (Shaddick et al., 2017).

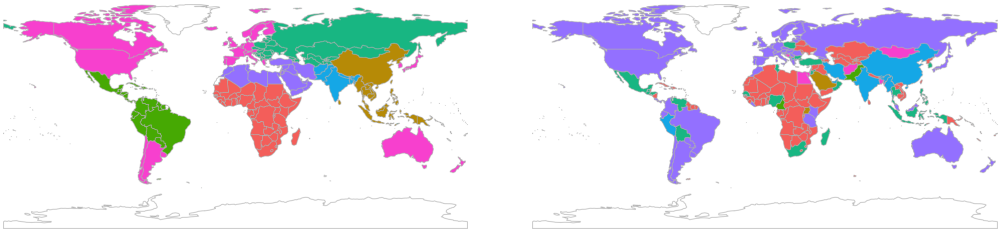
For our running example, we use the data from Shaddick et al. (2017), aggregated to the city level, to estimate ambient $PM_{2.5}$ concentration across the world. The statistical problem is that we only have direct measurements of $PM_{2.5}$ from a sparse network of 2980 ground monitors with heterogeneous spatial coverage (Figure 1a). This monitoring network has especially poor coverage across Africa, central Asia, and Russia.

In order to estimate the public health effect of $PM_{2.5}$, we need estimates of the $PM_{2.5}$ concentration at the same spatial resolution as population data. To obtain these estimates, we supplement the direct measurements with a high-resolution satellite data product that converts measurements of aerosol optical depth into estimates of $PM_{2.5}$. The hope is that we can use the ground monitor data to calibrate the approximate satellite measurements, and hence get estimates of $PM_{2.5}$ at the required spatial resolution.

The aim of this analysis is to build a predictive model of $PM_{2.5}$ with appropriately calibrated prediction intervals. We will not attempt a full analysis of this data, which was undertaken by Shaddick et al. (2017). Instead, we will focus on three simple, but plausible, models for the data in order to show how visualization can be used to help construct, sense-check, compute, and evaluate these models.

2. Exploratory data analysis goes beyond just plotting the data

An important aspect of formalizing the role of visualization in exploratory data analysis is to place it within the context of a particular statistical workflow. In particular, we argue that exploratory data analysis is more than simply plotting the data. Instead, we



(a) The WHO super-regions. The pink super-region corresponds to wealthy countries. The remaining regions are defined based on geographic contiguity.

(b) The super-regions found by clustering based on ground measurements of $PM_{2.5}$. Countries for which we have no ground monitor measurements are colored red.

Fig. 2: *World Health Organization super-regions and super-regions from clustering.*

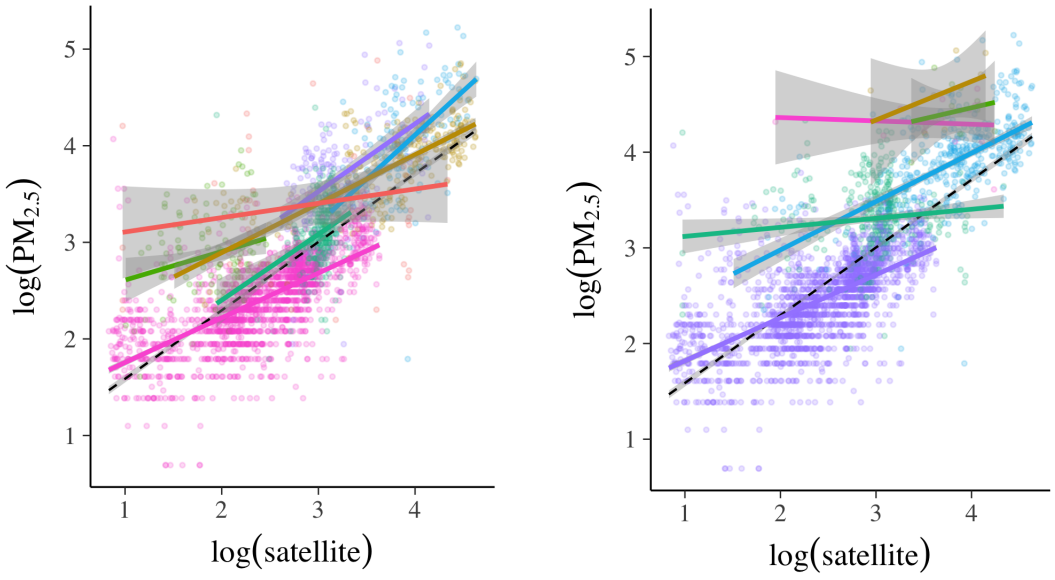
consider it a method to build a network of increasingly complex models that can capture the features and heterogeneities present in the data (Gelman, 2004).

This ground-up modeling strategy is particularly useful when the data that have been gathered are sparse or unbalanced, as the resulting network of models is built knowing the limitations of the design. A different strategy, which is common in machine learning, is to build a top-down model that throws all available information into a complicated non-parametric procedure. This works well for data that are a good representation the population of interest but can be prone to over-fitting or generalization error when used on sparse or unbalanced data. Using a purely predictive model to calibrate the satellite measurements would yield a fit that would be dominated by data in Western Europe and North America, which have very different air pollution profiles than most developing nations. With this in mind, we use the ground-up strategy to build a small network of three simple models for predicting $PM_{2.5}$ on a global scale.

The simplest predictive model that we can fit assumes that the satellite data product is a good predictor of the ground monitor data after a simple affine adjustment. In fact, this was the model used by the Global Burden of Disease project before the 2016 update (Forouzanfar et al., 2015). Figure 1b shows a straight line that fits the data on a log-log scale reasonably well ($R^2 \approx 0.6$). Discretization artifacts at the lower values of $PM_{2.5}$ are also clearly visible.

To improve the model, we need to think about possible sources of heterogeneity. For example, we know that developed and developing countries have different levels of industrialization and hence different air pollution. We also know that desert sand can be a large source of $PM_{2.5}$. If these differences are not appropriately captured by the satellite data product, fitting only a single regression line could leave us in danger of falling prey to Simpson's paradox (that a trend can reverse when data are grouped).

To expand out our network of models, we consider two possible groupings of countries. The WHO super-regions (Figure 2a) separate out rich countries and divide the remaining countries into six geographically contiguous regions. These regions have not been constructed with air pollution in mind, so we also constructed a different division based



(a) The same as Figure 1b, but also showing independent linear models fit within each WHO super-region.

(b) The same as (a), but the the linear models are fit within each of the cluster regions shown in Figure 2b.

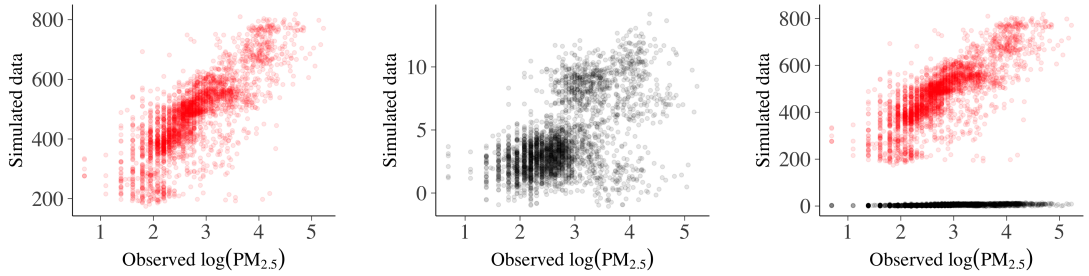
Fig. 3: *Graphics in model building: here, evidence that a single linear trend is insufficient.*

on a 6-component hierarchical clustering of ground monitor measurements of $PM_{2.5}$ (Figure 2b). The seventh region constructed this way is the collection of all countries for which we do not have ground monitor data.

When the trends for each of these regions are plotted individually (Figures 3a, 3b), it is clear that some ecological bias would creep into the analysis if we only used a single linear regression. We also see that some regions, particularly Sub-Saharan Africa (red in Figure 3a) and clusters 1 and 6 (pink and yellow in Figure 3b), do not have enough data to comprehensively nail down the linear trend. This suggests that some borrowing of strength through a multilevel model may be appropriate.

From this preliminary data analysis, we have constructed a network of three potential models. Model 1 is a simple linear regression. Model 2 is a multilevel model where observations are stratified by WHO super-region. Model 3 is a multilevel model where observations are stratified by *clustered* super-region.

These three models will be sufficient for demonstrating our proposed workflow, but this is a smaller network of models than we would use for a comprehensive analysis of the $PM_{2.5}$ data. Shaddick et al. (2017), for example, also consider smaller regions, country-level variation, and a spatial model for the varying coefficients. Further calibration covariates can also be included.



(a) Vague priors (b) Weakly informative priors (c) Comparison

Fig. 4: *Visualizing the prior predictive distribution. Panels (a) and (b) show realizations from the prior predictive distribution using vague priors and weakly informative priors, respectively. Simulated data are plotted on the y-axis and observed data on the x-axis. Because the simulations under the vague and weakly informative priors are so different, the y-axis scales used in panels (a) and (b) also differ dramatically. Panel (c) emphasizes the difference in the simulations by showing the red points from (a) and the black points from (b) plotted using the same y-axis.*

3. Fake data can be almost as valuable as real data for building your model

The exploratory data analysis resulted in a network of three models: one linear regression model and two different linear multilevel models. In order to fully specify these models, we need to specify prior distributions on all of the parameters. If we specify proper priors for all parameters in the model, a Bayesian model yields a joint prior distribution on parameters and data, and hence a prior marginal distribution for the data. That is, Bayesian models with proper priors are *generative models*. The main idea in this section is that we can visualize simulations from the prior marginal distribution of the data to assess the consistency of the chosen priors with domain knowledge.

The main advantage to assessing priors based on the prior marginal distribution for the data is that it reflects the interplay between the prior distribution on the parameters and the likelihood. This is a vital component of understanding how prior distributions actually work for a given problem (Gelman et al., 2017). It also explicitly reflects the idea that we can't fully understand the prior by fixing all but one parameter and assessing the effect of the unidimensional marginal prior. Instead, we need to assess the effect of the prior as a multivariate distribution.

The prior distribution over the data allows us to extend the concept of a weakly informative prior (Gelman et al., 2008) to be more aware of the role of the likelihood. In particular, we say that a prior leads to a *weakly informative joint prior data generating process* if draws from the prior data generating distribution $p(y)$ could represent any data set that could plausibly be observed. As with the standard concept of weakly informative priors, it's important that this prior predictive distribution for the data has at least some mass around extreme but plausible data sets. On the other hand, there should be no mass on completely implausible data sets. We recommend assessing how informative the prior distribution on the data is by generating a "flip book" of simulated datasets that can be used to investigate the variability and multivariate structure of the distribution.

To demonstrate the power of this approach, we return to the multilevel model for the $PM_{2.5}$ data. Mathematically, the model will look like $y_{ij} \sim \mathrm{N}(\beta_0 + \beta_{0j} + (\beta_1 + \beta_{1j})x_{ij}, \sigma^2)$, $\beta_{0j} \sim \mathrm{N}(0, \tau_0^2)$, $\beta_{1j} \sim \mathrm{N}(0, \tau_1^2)$, where y_{ij} is the logarithm of the observed $PM_{2.5}$, x_{ij} is the logarithm of the estimate from the satellite model, i ranges over the observations in each super-region, j ranges over the super-regions, and σ, τ_0, τ_1, β_0 and β_1 need prior distributions.

Consider some priors of the sort that are sometimes recommended as being vague: $\beta_k \sim N(0, 100)$, $\tau_k^2 \sim$ Inv-Gamma$(1, 100)$. The data generated using these priors and shown in Figure 4a are completely impossible for this application; note the y-axis limits and recall that the data are on the log scale. This is primarily because the vague priors don't actually respect our contextual knowledge.

We know that the satellite estimates are reasonably faithful representations of the $PM_{2.5}$ concentration, so a more sensible set of priors would be centered around models with intercept 0 and slope 1. An example of this would be $\beta_0 \sim N(0, 1)$, $\beta_1 \sim N(1, 1)$, $\tau_k \sim N_+(0, 1)$. Data generated by this model is shown in Figure 4b. While it is clear that this realization corresponds to a quite mis-calibrated satellite model (especially when we remember that we are working on the log scale), it is quite a bit more plausible than the model with vague priors.

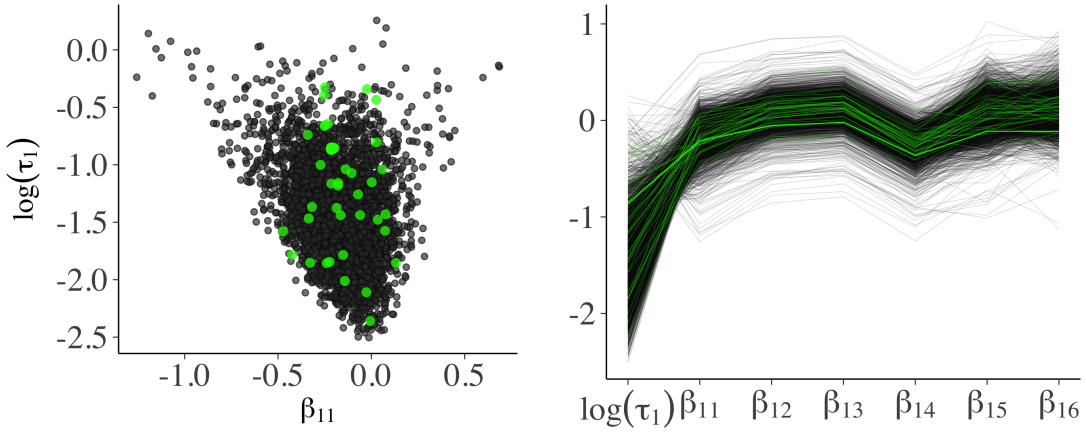
We argue that the tighter priors are still only weakly informative, in that the implied data generating process can still generate data that is much more extreme than we would expect from our domain knowledge. In fact, when repeating the simulation shown in Figure 4b many times we found that the data generated using these priors can produce data points with more than 22,000$\mu g m^{-3}$, which is a still a very high number in this context.

The prior predictive distribution is a powerful tool for understanding the structure of our model before we make a measurement, but its density evaluated at the measured data also plays the role of the marginal likelihood which is commonly used in model comparison. Unfortunately the utility of the the prior predictive distribution to evaluate the model does not extend to utility in selecting between models. For further discussion see Gelman et al. (2017).

4. Graphical Markov chain Monte Carlo diagnostics: moving beyond trace plots

Constructing a network of models is only the first step in the Bayesian workflow. Our next job is to fit them. Once again, visualizations can be a key tool in doing this well. Traditionally, Markov chain Monte Carlo (MCMC) diagnostic plots consist of trace plots and autocorrelation functions. We find these plots can be helpful to understand problems that have been caught by numerical summaries such as the potential scale reduction factor $\widehat{R}$ (Stan Development Team, 2017b, Section 30.3), but they are not always needed as part of workflow in the many settings where chains mix well.

For general MCMC methods it is difficult to do any better than between/within summary comparisons, following up with trace plots as needed. But if we restrict our attention to Hamiltonian Monte Carlo (HMC) and its variants, we can get much more detailed information about the performance of the Markov chain (Betancourt, 2017). We know that the success of HMC requires that the geometry of the set containing the bulk



(a) For Model 3, a bivariate plot of the log standard deviation of the cluster-level slopes (y-axis) against the slope for the first cluster (x-axis). The green dots indicate starting points of divergent transitions. This plot can be made using `mcmc_scatter` in `bayesplot`.

(b) For Model 3, a parallel coordinates plot showing the cluster-level slope parameters and their log standard deviation $\log \tau_1$. The green lines indicate starting points of divergent transitions. This plot can be made using `mcmc_parcoord` in `bayesplot`.

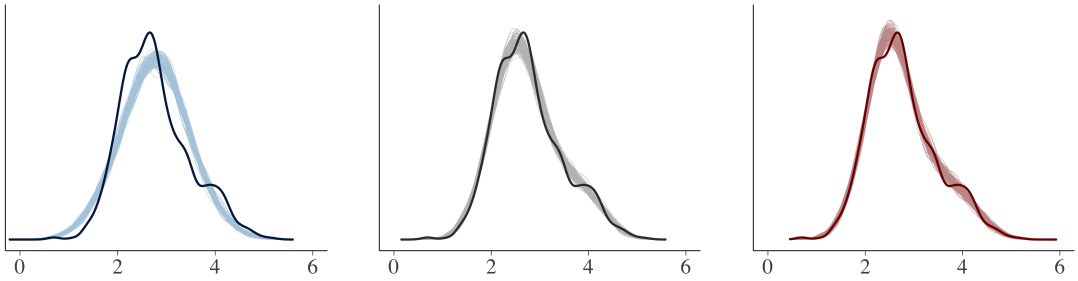
Fig. 5: *Several different diagnostic plots for Hamiltonian Monte Carlo. Models were fit using the RStan interface to Stan 2.17 (Stan Development Team, 2017a).*

of the posterior probability mass (which we call the typical set) is fairly smooth. It is not possible to check this condition mathematically for most models, but it can be checked numerically. It turns out that if the geometry of the typical set is non-smooth, the path taken by leap-frog integrator that defines the HMC proposal will rapidly diverge from the energy conserving trajectory.

Diagnosing divergent numerical trajectories precisely is difficult, but it is straightforward to identify these divergences heuristically by checking if the error in the Hamiltonian crosses a large threshold. Occasionally this heuristic falsely flags stable trajectories as divergent, but we can identify these false positives visually by checking if the samples generated from divergent trajectories are distributed in the same way as the non-divergent trajectories. Combining this simple heuristic with visualization greatly increases its value.

Visually, a concentration of divergences in small neighborhoods of parameter space, however, indicates a region of high curvature in the posterior that obstructs exploration. These neighborhoods will also impede any MCMC method based on local information, but to our knowledge only HMC has enough mathematical structure to be able to reliably diagnose these features. Hence, when we are using HMC for our inference, we can use visualization to not just assess the convergence of the MCMC method, but also to understand the geometry of the posterior.

There are several plots that we have found useful for diagnosing troublesome areas of the parameter space, in particular bivariate scatterplots that mark the divergent transitions (Figure 11a), and parallel coordinate plots (Figure 11b). These visualizations are sensitive enough to differentiate between models with a non-smooth typical set and



(a) Model 1 (b) Model 2 (c) Model 3

Fig. 6: *Kernel density estimate of the observed dataset y (dark curve), with density estimates for 100 simulated datasets* y_{rep} *drawn from the posterior predictive distribution (thin, lighter lines). These plots can be produced using* `ppc_dens_overlay` *in the* `bayesplot` *package.*

models where the heuristic has given a false positive. This makes them an indispensable tool for understanding the behavior of a HMC when applied to a particular target distribution.

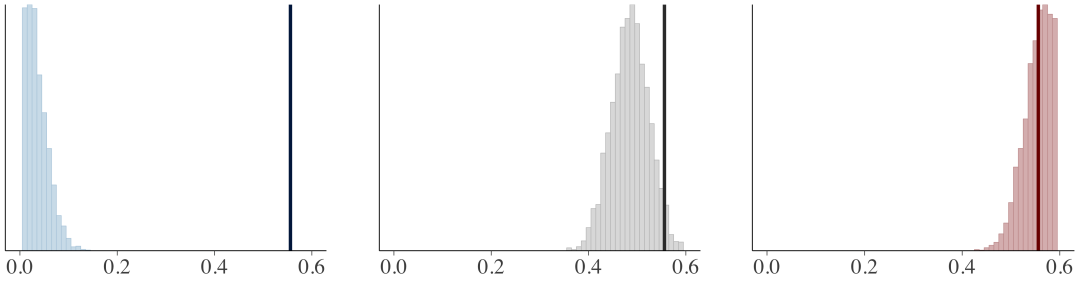
If HMC were struggling to fit Model 3, the divergences would be clustered in the parameter space. Examining the bivariate scatterplots (Figure 11a), there is no obvious pattern to the divergences. Similarly, the parallel coordinate plot (Figure 11b) does not show any particular structure. This indicates that the divergences that are found are most likely false positives. For contrast, the supplementary material contains the same plots for a model where HMC fails to compute a reliable answer. In this case, the clustering of divergences is pronounced and the parallel coordinate plot clearly indicates that all of the divergent trajectories have the same structure.

5. How did we do? Posterior predictive checks are vital for model evaluation

The idea behind posterior predictive checking is simple: if a model is a good fit we should be able to use it to generate data that resemble the data we observed. This is similar in spirit to the prior checks considered in Section 3, except now we have a data-informed data generating model. This means we can be much more stringent in our comparisons. Ideally, we would compare the model predictions to an independent test data set, but this is not always feasible. However, we can still do some checking and predictive performance assessments using the data we already have.

To generate the data used for posterior predictive checks (PPCs) we simulate from the posterior predictive distribution $p(\tilde{y} \mid y) = \int p(\tilde{y} \mid \theta) p(\theta \mid y)\, d\theta$, where y is our current data, $\tilde{y}$ is our new data to be predicted, and θ are our model parameters. Posterior predictive checking is mostly qualitative. By looking at some important features of the data and the replicated data, which were not explicitly included in the model, we may find a need to extend or modify the model.

For each of the three models, Figure 6 shows the distributions of many replicated datasets drawn from the posterior predictive distribution (thin light lines) compared to



(a) Model 1 (b) Model 2 (c) Model 3

Fig. 7: *Histograms of statistics* $\text{skew}(y_{\text{rep}})$ *computed from 4000 draws from the posterior predictive distribution. The dark vertical line is computed from the observed data. These plots can be produced using* `ppc_stat` *in the* `bayesplot` *package.*

the empirical distribution of the observed outcome (thick dark line). From these plots it is evident that the multilevel models (2 and 3) are able to simulate new data that is more similar to the observed $\log(\text{PM})_{2.5}$ values than the model without any hierarchical structure (Model 1).

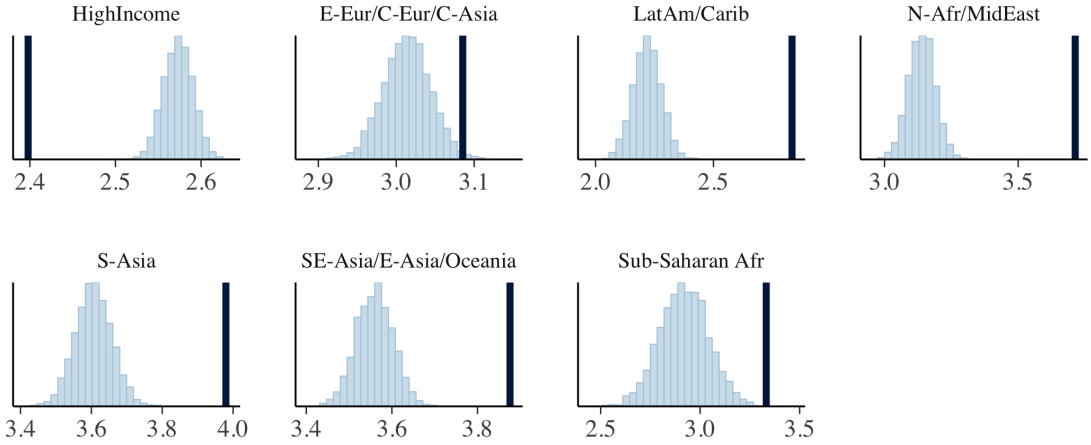
Posterior predictive checking makes use of the data twice, once for the fitting and once for the checking. Therefore it is a good idea to choose statistics that are orthogonal to the model parameters. If the test statistic is related to one of the model parameters, e.g., if the mean statistic is used for a Gaussian model with a location parameter, the posterior predictive checks may be less able to detect conflicts between the data and the model. Our running example uses a Gaussian model so in Figure 7 we investigate how well the posterior predictive distribution captures skewness. Model 3, which used data-adapted regions, is best at capturing the observed skewness, while Model 2 does an ok job and the linear regression (Model 1) totally fails.

We can also perform similar checks within levels of a grouping variable. For example, in Figure 8 we split both the outcome and posterior predictive distribution according to region and check the median values. The two hierarchical models give a better fit to the data at the group level, which in this case is unsurprising.

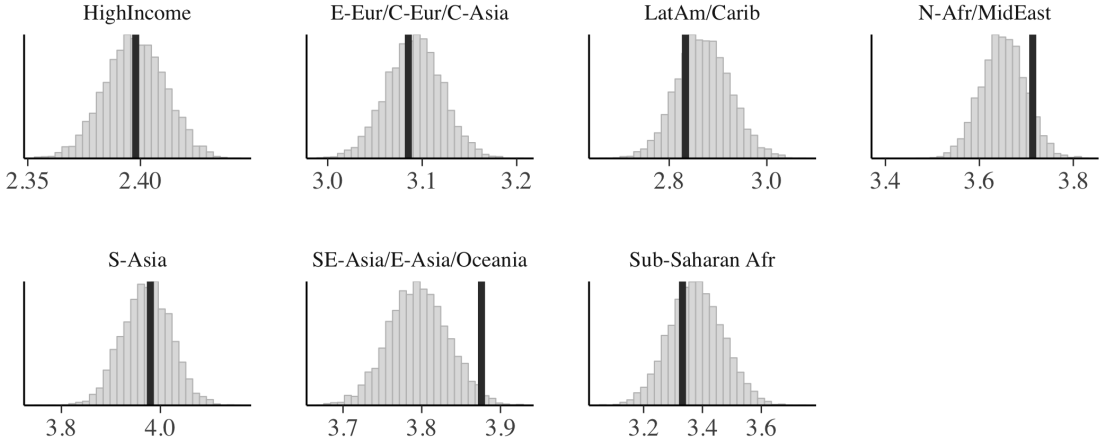
In cross-validation, double use of data is partially avoided and test statistics can be better calibrated. When performing leave-one-out (LOO) cross-validation we usually work with univariate posterior predictive distributions, and thus we can't examine properties of the joint predictive distribution. To specifically check that predictions are calibrated, the usual test is to look at the leave-one-out cross-validation predictive cumulative density function values, which are asymptotically uniform (for continuous data) if the model is calibrated (Gelfand et al., 1992; Gelman et al., 2013).

The plots shown in Figure 9 compare the density of the computed LOO-PITs (thick dark line) versus 100 simulated datasets from a standard uniform distribution (thin light lines). We can see that, although there is some clear miscalibration in all cases, the hierarchical models are an improvement over the single-level model.

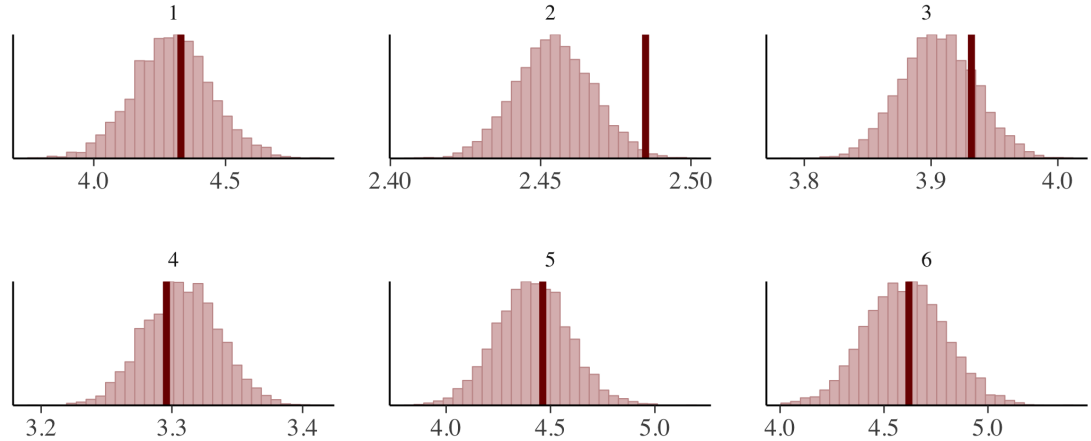
The shape of the miscalibration in Figure 9 is also meaningful. The frown shapes exhibited by Models 2 and 3 indicate that the univariate predictive distributions are too



(a) Model 1

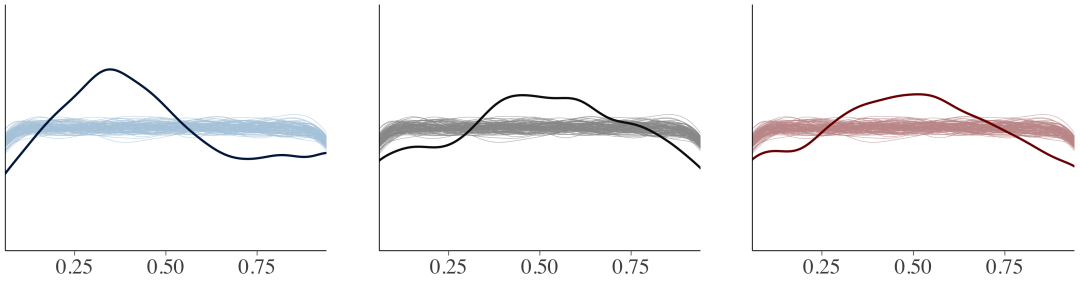


(b) Model 2



(c) Model 3

Fig. 8: *Checking posterior predictive test statistics, in this case the medians, within region. The vertical lines are the observed medians. The facets are labelled by number in panel (c) because they represent groups found by the clustering algorithm rather than actual super-regions. These grouped plots can be made using* `ppc_stat_grouped` *in the* `bayesplot` *package.*



(a) Model 1 (b) Model 2 (c) Model 3

Fig. 9: *Graphical check of leave-one-out cross-validated probability integral transform (LOO-PIT). The thin lines represent simulations from the standard uniform distribution and the thick dark line in each plot is the density of the computed LOO-PITs. Similar plots can be made using* `ppc_dens_overlay` *and* `ppc_loo_pit` *in the* `bayesplot` *package. The downwards slope near zero and one on the "uniform" histograms is an edge effect due to the density estimator used and can be safely discounted.*

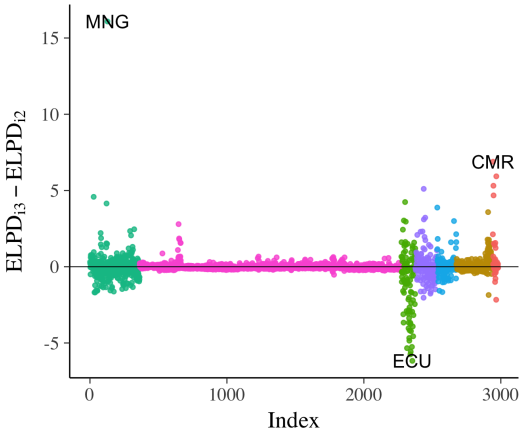
broad compared to the data, which suggests that further modeling will be necessary to accurately reflect the uncertainty. One possibility would be to further sub-divide the super-regions to better capture within-region variability (Shaddick et al., 2017).

6. Pointwise plots for predictive model comparison

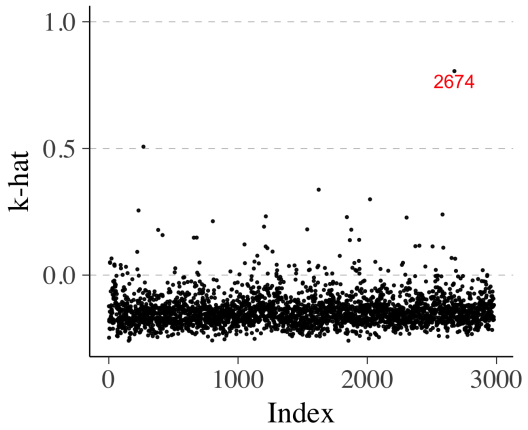
Visual posterior predictive checks are also useful for identifying unusual points in the data. Unusual data points come in two flavors: outliers and points with high leverage. In this section, we show that visualization can be useful for identifying both types of data point. Examining these unusual observations is a critical part of any statistical workflow, as these observations give hints as to how the model may need to be modified. For example, they may indicate the model should use non-linear instead of linear regression, or that the observation error should be modeled with a heavier tailed distribution.

The main tool in this section is the one-dimensional cross-validated leave-one-out (LOO) predictive distribution $p(y_i \mid y_{-i})$. Gelfand et al. (1992) suggested examining the LOO log-predictive density values (they called them conditional predictive ordinates) to find observations that are difficult to predict. This idea can be extended to model comparison by looking at which model best captures each left out data point. Figure 10a shows the difference between the expected log predictive densities (ELPD) for the individual data points estimated using Pareto-smoothed importance sampling (PSIS-LOO, Vehtari et al. (2017b,c)). Model 3 appears to be slightly better than Model 2, especially for difficult observations like the station in Mongolia.

In addition to looking at the individual LOO log-predictive densities, it is useful to look at how influential each observation is. Some of the data points may be difficult to predict but not necessarily influential, that is, the predictive distribution does not change much when they are left out. One way to look at the influence is to look at the difference between full data log-posterior predictive density and the LOO log-predictive density.



(a) The difference in pointwise ELPD values obtained from PSIS-LOO for Model 3 compared to Model 2 colored by the WHO cluster (see Figure 1b for the key). Positive values indicate Model 3 outperformed Model 2.



(b) The $\hat{k}$ diagnostics from PSIS-LOO for Model 2. The 2674th data point (the only data point from Mongolia) is highlighted by the $\hat{k}$ diagnostic as being influential on the posterior.

Fig. 10: *Model comparisons using leave-one-out (LOO) cross-validation.*

We recommend computing the LOO log-predictive densities using PSIS-LOO as implemented in the `loo` package (Vehtari et al., 2017a). A key advantage of using PSIS-LOO to compute the LOO densities is that it automatically computes an empirical estimate of how similar the full-data predictive distribution is to the LOO predictive distribution for each left out point. Specifically, it computes an empirical estimate $\hat{k}$ of $k = \inf\left\{k' > 0 : D_{\frac{1}{k'}}\left(p||q\right) < \infty\right\}$, where $D_\alpha\left(p||q\right) = \frac{1}{\alpha-1}\log\int_\Theta p(\theta)^\alpha q(\theta)^{1-\alpha}d\theta$ is the α-Rényi divergence (Yao et al., 2018). If the jth LOO predictive distribution has a large $\hat{k}$ value when used as a proposal distribution for the full-data predictive distribution, it suggests that y_j is a highly inferential observation.

Figure 10b shows the $\hat{k}$ diagnostics from PSIS-LOO for our Model 2. The 2674th data point is highlighted by the $\hat{k}$ diagnostic as being influential on the posterior. If we examine the data we find that this point is the only observation from Mongolia and corresponds to a measurement $(x, y) = (\log\left(\text{satellite}\right), \log\left(\text{PM}_{2.5}\right)) = (1.95, 4.32)$, which would look like an outlier if highlighted in the scatterplot in Figure 1b. By contrast, under Model 3 the $\hat{k}$ value for the Mongolian observation is significantly lower ($\hat{k} \approx 0.5$) indicating that that point is better resolved in Model 3.

7. Discussion

Visualization is probably the most important tool in an applied statistician's toolbox and is an important complement to quantitative statistical procedures. In this paper, we've demonstrated that it can be used as part of a strategy to compare models, to identify ways in which a model fails to fit, to check how well our computational methods have

resolved the model, to understand the model well enough to be able to set priors, and to iteratively improve the model.

The last of these tasks is a little bit controversial as using the measured data to guide model building raises the concern that the resulting model will generalize poorly to new datasets. A different objection to using the data twice (or even more) comes from ideas around hypothesis testing and unbiased estimation, but we are of the opinion that the danger of overfitting the data is much more concerning (Gelman and Loken, 2014).

In the visual workflow we've outlined in this paper, we have used the data to improve the model in two places. In Section 3 we proposed prior predictive checks with the recommendation that the data generating mechanism should be broader than the distribution of the observed data in line with the principle of weakly informative priors. In Section 5 we recommended undertaking careful calibration checks as well as checks based on summary statistics, and then updating the model accordingly to cover the deficiencies exposed by this procedure. In both of these cases, we have made recommendations that aim to reduce the danger. For the prior predictive checks, we recommend not cleaving too closely to the observed data and instead aiming for a prior data generating process that can produce plausible data sets, not necessarily ones that are indistinguishable from observed data. For the posterior predictive checks, we ameliorate the concerns by checking carefully for influential measurements and proposing that model extensions be weakly informative extensions that are still centered on the previous model (Simpson et al., 2017).

Regardless of concerns we have about using the data twice, the workflow that we have described in this paper (perhaps without the stringent prior and posterior predictive checks) is common in applied statistics. As academic statisticians, we have a duty to understand the consequences of this workflow and offer concrete suggestions to make the practice of applied statistics more robust.

Acknowledgements

The authors thank Gavin Shaddick and Matthew Thomas for their help with the $PM_{2.5}$ example, Ghazal Fazelnia for finding an error in our map of ground monitor locations, Ari Hartikainen for suggesting the parallel coordinates plot, and the Sloan Foundation, Columbia University, U.S. National Science Foundation, Institute for Education Sciences, and Office of Naval Research for financial support.

References

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. arXiv preprint arxiv:1701.02434.

Betancourt, M. and M. Girolami (2015). Hamiltonian Monte Carlo for hierarchical models. In S. K. Upadhyay, U. Singh, D. K. Dey, and A. Loganathan (Eds.), *Current Trends in Bayesian Methodology with Applications*, pp. 79–101. Chapman & Hall. arXiv:1312.0906.

Forouzanfar, M. H., L. Alexander, H. R. Anderson, V. F. Bachman, S. Biryukov, M. Brauer, R. Burnett, D. Casey, M. M. Coates, A. Cohen, et al. (2015). Global, re-

gional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet 386*(10010), 2287–2323.

Gabry, J. (2017). bayesplot: Plotting for Bayesian models. R package version 1.3.0, `http://mc-stan.org/bayesplot`.

Gelfand, A. E., D. K. Dey, and H. Chang (1992). Model determination using predictive distributions with implementation via sampling-based methods (with discussion). In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics 4*, pp. 147–167. Oxford University Press.

Gelman, A. (2004). Exploratory data analysis for complex models. *Journal of Computational and Graphical Statistics 13*(4), 755–779.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian Data Analysis* (Third ed.). Chapman & Hall/CRC. Chapter 6, Section "Marginal predictive checks".

Gelman, A., A. Jakulin, M. G. Pittau, Y.-S. Su, et al. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics 2*(4), 1360–1383.

Gelman, A. and E. Loken (2014). The statistical crisis in science: Data-dependent analysis-a "garden of forking paths"-explains why many statistically significant comparisons don't hold up. *American Scientist 102*(6), 460.

Gelman, A., D. Simpson, and M. Betancourt (2017). The prior can generally only be understood in the context of the likelihood. *arXiv preprint arXiv:1708.07487*.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics 6*, 377–401.

Shaddick, G., M. L. Thomas, A. Green, M. Brauer, A. Donkelaar, R. Burnett, H. H. Chang, A. Cohen, R. V. Dingenen, C. Dora, et al. (2017). Data integration model for air quality: a hierarchical approach to the global estimation of exposures to ambient air pollution. *Journal of the Royal Statistical Society: Series C (Applied Statistics) Available online 13 June 2017*. arXiv:1609.00141.

Simpson, D., H. Rue, A. Riebler, T. G. Martins, S. H. Sørbye, et al. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science 32*(1), 1–28.

Stan Development Team (2017a). RStan: the R interface to Stan, version 2.16.1. `http://mc-stan.org`.

Stan Development Team (2017b). *Stan Modeling Language User's Guide and Reference Manual, Version 2.16.0*. `http://mc-stan.org`.

Vehtari, A., A. Gelman, and J. Gabry (2017a). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. R package version 1.0.0, `http://mc-stan.org/loo`.

Vehtari, A., A. Gelman, and J. Gabry (2017b). Pareto smoothed importance sampling. arXiv preprint arXiv:1507.02646.

Vehtari, A., A. Gelman, and J. Gabry (2017c). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing 27*(5), 1413–1432. arXiv:1507.04544.

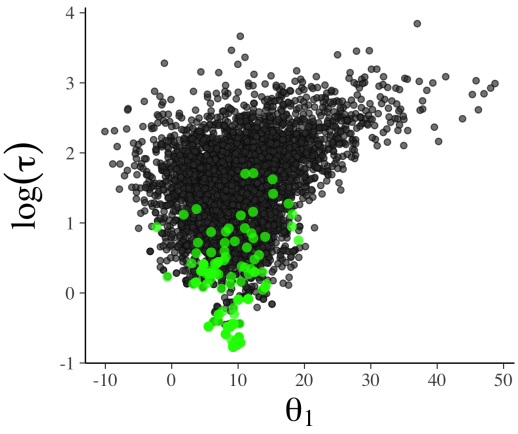
Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Yao, Y., A. Vehtari, D. Simpson, and A. Gelman (2018). Yes, but did it work?: Evaluating variational inference. *arXiv preprint arXiv:1802.02538* .

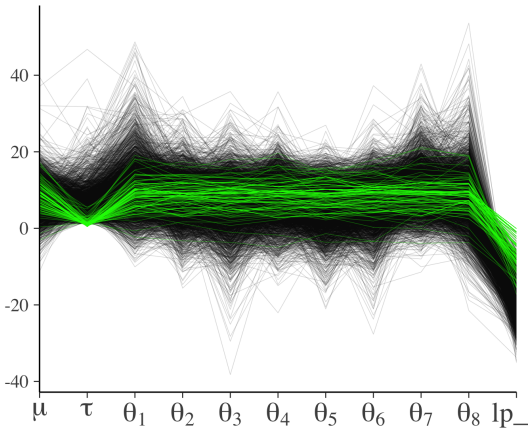
Supplementary Material: The 8-schools problem and the visualization of divergent trajectories

Consider the hierarchical 8-schools problem outlined in (Rubin, 1981; Gelman et al., 2013). Figure 11a shows a scatterplot of the log standard deviation of the school-specific parameters (τ, y-axis) against the parameter representing the mean for the first school (θ_1, x-axis). The starting points of divergent transitions, shown in green, concentrate in a particular region which is evidence of a geometric pathology in parameter space. Figure 11b gives a different perspective on the divergences. It is a parallel coordinates plot including *all* parameters from the 8-schools example with divergent iterations also highlighted in green. We can see in both the bivariate plot and the parallel coordinates plot that the divergences tend to occur when the hierarchical standard deviation τ goes to 0 and the values of the θ_j's are nearly constant. These problems can be fixed by re-parameterization (Betancourt and Girolami, 2015).

Now that we know precisely what part of the parameter space is causing problems, we can fix it. Funnels in the parameter space can be resolved through a reparameterization that fattens out the problem area. The standard tool for fixing funnels caused by hierarchical models is moving to a non-centered parameterization, where the narrowest coordinate is made a priori independent of the other coordinates in the funnel (Betancourt and Girolami, 2015). This will typically fatten out the funnel and remove the cluster of divergences.



(a) A bivariate plot of the log standard deviation of school-level effects ($\log(\tau)$, y-axis) against the mean for the first school (θ_1, x-axis) for the 8-schools problem. The green dots indicate starting points of divergent transitions. The pile up of divergences in a corner of the samples (in this case the neck of the funnel shape) strongly indicates that there is a problem with this part of the parameter space. This plot can be made using `mcmc_scatter` in `bayesplot`.



(b) Parallel coordinates plot for the 8-schools problem showing the school-specific parameters $(\theta_1, \ldots, \theta_8)$ and their prior mean and standard deviation (μ, τ). The green lines indicate the starting points of divergent transitions. In this case it is clear that all of the divergent paths have a small value of τ, which results in little variability in the θ_j's (the green lines are flat). This plot can be made using `mcmc_parcoord` in `bayesplot`.

Fig. 11: *Several different diagnostic plots for Hamiltonian Monte Carlo. Models were fit using the RStan interface to Stan 2.17 (Stan Development Team, 2017a).*