Generalised additive mixed models for large datasets: modeling ruminal temperature sensor data from dairy cows

Biometrics in the Bay of Islands, New Zealand

Chanatda Somchit
AgResearch Ltd., New Zealand

December 1, 2023

Outline

Generalised additive models (GAMs)

- GAM in `gam`, GAM in `mgcv`, and GAM in `VGAM`
- GAMs for large datasets
- GAMs for gigadata

Modeling ruminal temperature sensor data from dairy cows

- Heat stress in dairy cows
- Data
- Generalised additive mixed model fitting

Discussion: the use of the more modern GAMs for large datasets in a modern agriculture application

GAM in `gam`, GAM in `mgcv`, and GAM in `VGAM`

Generalised additive models (GAMs) are a nonparametric extension of generalised linear models.

- The GAMs basic-structure can be written as:

$$g(\mu) = \eta(\boldsymbol{x}) = \beta_1 x_1 + f_2(x_2) + \cdots + f_p(x_p). \tag{1}$$

In `R`, there are three main packages for GAM:

- The `gam` package [Hastie and Tibshirani, 1986, 1990] using a back-fitting algorithm.
- The `mgcv` package [Wood, 2000]: Mixed GAM Computation Vehicle using a penalised likelihood-based method.
- The `VGAM` package [Yee and Wild, 1996]: Vector Generalised Additive Models (VGAMs) extended the class of GAMs to include classes of multivariate regression models.

GAM in `mgcv`

The GAM [Wood, 2000] has a structure in the form of:

$$\eta_i \;=\; \mathbf{A}_i\boldsymbol{\theta} + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \ldots \qquad (2)$$

- Defining as an additive model, an interaction smooth effect, and smooth-factor interactions.

The main advantages of GAM by Wood [2000]:

- Allowing for generalised additive mixed modelling (GAMM) including spatial and temporal correlations as well as nested data and various heterogeneity structures.
- Covering a wide range of distributions beyond the usual GLM exponential family.

Large datasets

- Tens of thousands to millions of observations are now common.
- Current fitting methods for GAMs are reasonably efficient with up to a few tens of thousands of observations.
- Hundreds of thousands or millions of observations: GAMs tend to become too memory intensive.

GAMs for large datasets

Wood et al. [2015] developed practical generalised additive model fitting methods for large data sets.

- Using simple strategies (reducing storage-use inefficiencies) for updating a model matrix factorisation: requiring only sub-blocks of the model matrix to be computed at any one time.
- The methods are implemented in `R` package `mgcv` as function `bam`.

GAMs for gigadata

Much larger datasets: Wood et al. [2017] developed scalable methods for fitting GAMs with of the order of 10^4 coefficients to up to 10^8 data.

- Based on three innovations:
  1. the development of a fitting method which required only basic easily parallelised matrix computations and a pivoted Cholesky decompostion;
  2. the use of a scalable parallel block pivoted Cholesky algorithm;
  3. an efficient approach to model matrix storage and computations with the model matrix, using discretised covariates.
- The methods are implemented in the `bam` function via arguments `discrete` and `nthreads`.

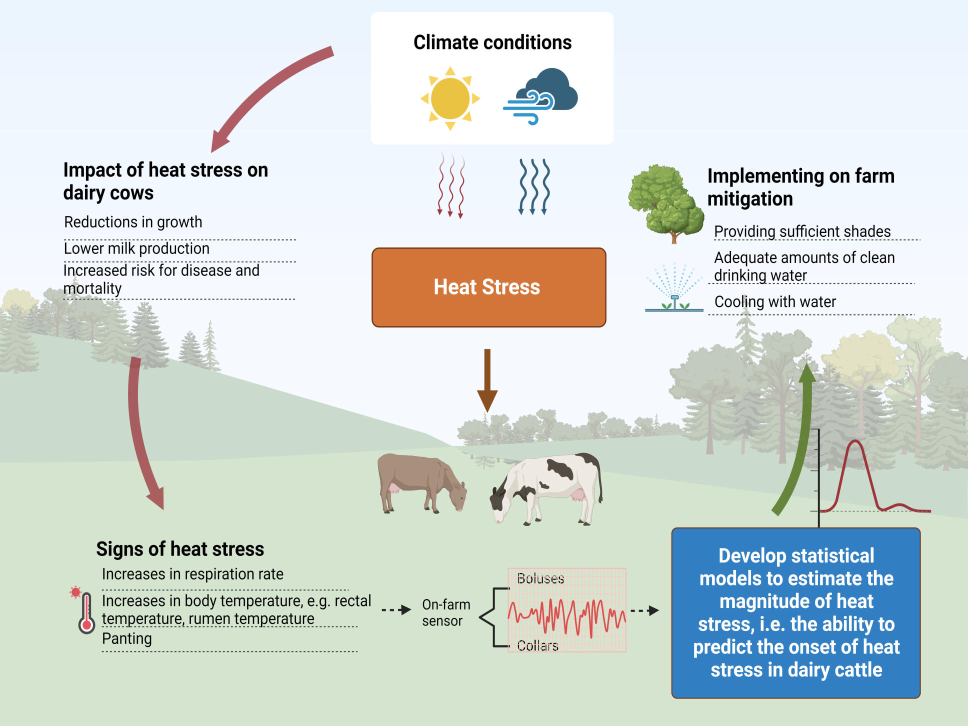
Heat stress in dairy cows

The research program: The New Zealand Bioeconomy In the Digital Age (NZBIDA): An animal-centric dairy industry enabled by digital technology.

https://www.tdm.it/en/project/afi-collar/

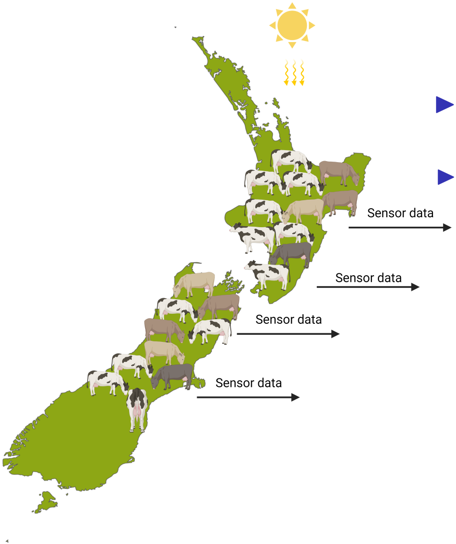
- Uses existing on-farm sensor technologies (e.g., boluses and collars) for identification of lead and lag indicators of heat stress (e.g., rumen temperature and panting).

agresearch
āta mātai, mātai whetū



Climate conditions
Impact of heat stress on dairy cows
Reductions in growth
Lower milk production
Increased risk for disease and mortality
Heat Stress
Implementing on farm mitigation
Providing sufficient shades
Adequate amounts of clean drinking water
Cooling with water
Signs of heat stress
Increases in respiration rate
Increases in body temperature, e.g. rectal temperature, rumen temperature
Panting
On-farm sensor
Boluses
Collars
Develop statistical models to estimate the magnitude of heat stress, i.e. the ability to predict the onset of heat stress in dairy cattle

Data



- Rumen temperature (°C) was recorded at 10-min intervals.
- Climate variables were recorded at 15-min intervals:
  - Temperature (°C),
  - Relative humidity (%),
  - Solar radiation (MJ/m^2),
  - Wind speed (m/s),
  - Cumulative solar radiation (MJ/m^2),
  - Temperature-humidity index (THI).

Sensor Data

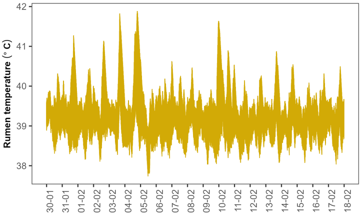


Figure 1: 10-min rumen temperature measurements during summer 2023 for 1 farm (traces from 80 cows).

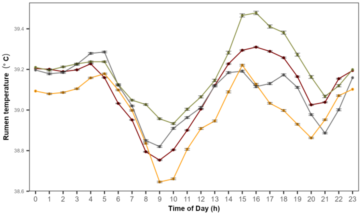


Figure 2: Hourly averages across all cows for each of 4 farms.

agresearch
āta mātai, mātai whetū

Climate Data

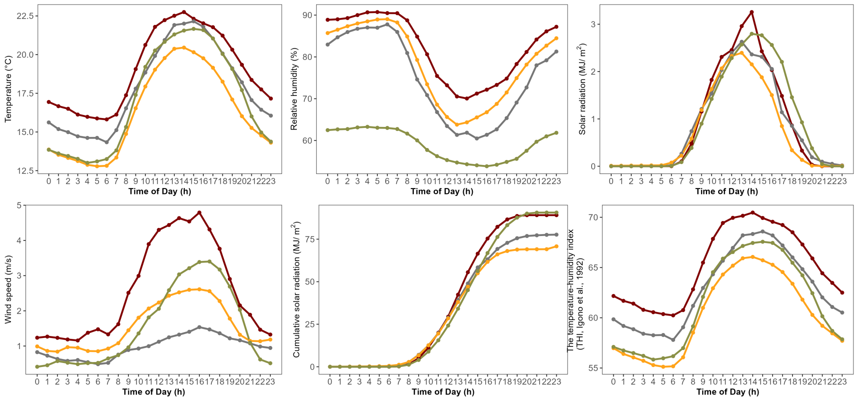


Figure 3: Average temperature, relative humidity, solar radiation, wind speed, cumulative solar radiation, and temperature-humidity index over time of day for each of 4 farms.

agresearch
āta mātai, mātai whetū

Generalised additive mixed model fitting

Best-fit GAMM for the rumen temperature as a response variable for the sensor data obtained during summer 2023.

1. Smooth 'main effects'

$$\begin{aligned} \text{Rumen1}_{ijk} &= \beta_1 + f_1(\text{ToD}_{ijk}) + f_2(\text{Prev. H THI}_{ijk}) + f_3(\text{Prev. H WS}_{ijk}) \\ &+ f_4(\text{Prev. H CuSR}_{ijk}) + a_i + b_{ij} + \varepsilon_{ijk} \end{aligned}$$

2. Smooth 'interactions'

$$\begin{aligned} \text{Rumen2}_{ijk} &= \beta_1 + f_1(\text{ToD}_{ijk}) + f_2(\text{Prev. H THI}_{ijk}, \text{Prev. H WS}_{ijk}) \\ &+ f_3(\text{Prev. H THI}_{ijk}, \text{Prev. H CuSR}_{ijk}) \\ &+ f_4(\text{Prev. H WS}_{ijk}, \text{Prev. H CuSR}_{ijk}) \\ &+ a_i + b_{ij} + \varepsilon_{ijk} \end{aligned}$$

Where, Prev. H = previous hr.

agresearch
āta mātai, mātai whetū

Generalised additive mixed model fitting

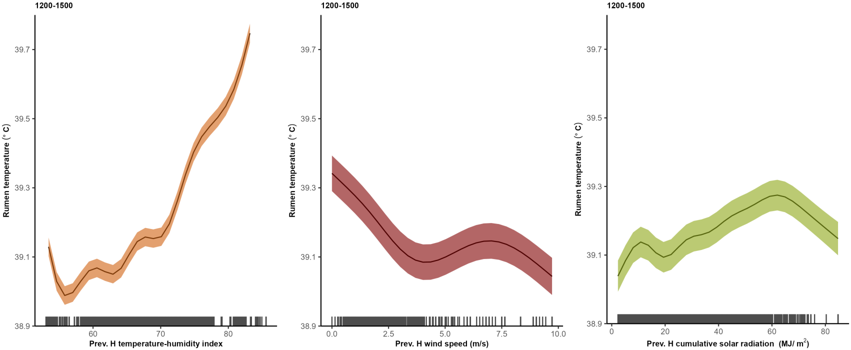


Figure 4: Estimated previous hour THI, wind speed, and cumulative solar radiation effects from the GAMM with 95% confidence intervals.

Generalised additive mixed model fitting

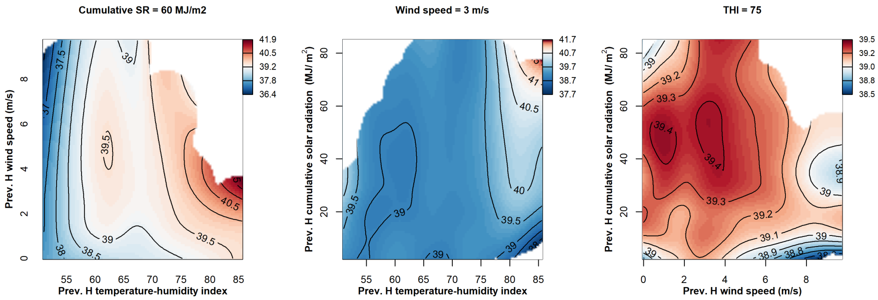


Figure 5: Contour plots of rumen temperature (°C) on combined effects of (i) previous hour THI and wind speed, (ii) previous hour THI and cumulative solar radiation, and (iii) previous hour wind speed and cumulative solar radiation from the GAMM.

agresearch
āta mātai, mātai whetū

Discussion

- Wood et al. [2015, 2017] developed methods allowing estimation of GAMs to be feasible for large and much larger data sets.
- Clearly, GAMs are powerful tools used to identify or predict the onset of heat stress in dairy cattle (on-farm sensor data), which is a key element of successful on farm mitigation.

Thanks

- Kirsty Verhoek and DairyNZ team
- AgResearch team
- Fonterra

References

T. J. Hastie and R. J. Tibshirani. Generalized additive models. *Statistical science*, pages 297–310, 1986.

T. J. Hastie and R. J. Tibshirani. *Generalized additive models*. Chapman and Hall, London, 1990.

S. N. Wood. Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62:413–428, 2000.

S. N. Wood, Y. Goude, and S. Shaw. Generalized additive models for large data sets. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 64(1):139–155, 2015.

S. N. Wood, Z. Li, G. Shaddick, and N. H. Augustin. Generalized additive models for gigadata: modeling the U.K. black smoke network daily data. *Journal of the American Statistical Association*, 112(519):1199–1210, 2017.

T. W. Yee and C. J. Wild. Vector generalized additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58:481–493, 1996.

agresearch
āta mātai, mātai whetū