Application of COWBASE for analysis of activity and milk production changes in function of THI in different countries.

# Introduction to the case

* Relevance: moderate climate, etc
* Individual variation: why do we need to know how cows react to thi
  + Physiological processes
  + Welfare and health
  + Long-short term consequences
  + Breeding and monitoring
  + Future-proof dairy farming
  + Tailor management to cows and vice versa
* Value of (simultaneous) evaluation of multiple parameters
* Our study:
  + Selection of multiple farms
  + Different countries and environments/climatic conditions
  + Longer period of time, with data that spans multiple years and so also different conditions (warmer/colder summers)

# Materials and methods

To demonstrate the value of CowBase, we performed an simple analysis to explore farm and individual cow variation of daily milk yield (**DMY**) and activity (step count, **ACT**) in response to changing outdoor climatic conditions. CowBase allows to rapidly access data that are ready-to-analyze, are consistent in format and can easily be combined based on farm-animal-lactation or datetime identifiers.

## Data description and preprocessing

We extracted datasets with activity, milk production, cow ancillary information and weather data from 1 Italian, 3 Belgian and 2 Dutch farms with an Lely automatic milking system from CowBase using module **xxx**. We selected years of data for which the activity data had stable statistical properties indicating continuity in hardware and software settings. An overview of the included farms is given in Table **T1**. The farms all had an intensive production system with cows kept indoors, year-round calving and fed with both roughage and concentrates on a production-level basis. All analysis were performed in Python 3.10, using *pandas*, *numpy*, *matplotlib*, *seaborn* and *statsmodels* packages. Version specifics and environment requirements can be found in the “*requirements.txt*” file of the repository at [**link**].

**Table T1**. overview of farm data:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Farm | Country/region | Data period | no. of cows | | No. of lactations | | 305d production level [kg] |
| activity | milk yield | activity | milk yield |
| 1 | BE – west (seaside) | 01/2011-12/2019 | 255 | 358 | 401 | 653 | 8696±1613 |
| 2 | NL – east (inland) | 01/2014-12/2017 | 677 | 714 | 1567 | 1726 | 10118±1677 |
| 3 | NL – west | 01/2014-12/2017 | 188 | 242 | 354 | 505 | 9170±1783 |
| 4 | BE – east | 01/2017-12/2022 | 281 | 310 | 624 | 732 | 10461±1785 |
| 5 | BE – central | 01/2016-12/2019 | 147 | 161 | 257 | 345 | 11072±1745 |
| 6 | IT - ?? | 01/2013-12/2020 | 290 | 371 | 482 | 762 | 9963±2053 |

### Daily milk yield

Daily milk yield (DMY) was calculated as the summed milk yield of individual milkings per 24 hours. To account for unequal milking intervals and the varying number of milkings per day, the percentage of milk of the first milking of each day produced on the previous day, was partitioned across both days according to the milking interval and milking time. For example, for a 4 AM milking with a milking interval from the previous one of 10 hours, 40% of the milk yield was assigned to the current and 60% to the previous day. Next, we selected lactations that (1) had at least 65 days of milk yield available from the onset of the lactation (DIM at start of the dataset below 5 days in lactation) and (2) that did not contain gaps longer than 5 successive days. Daily milk yield produced later than 400 days in lactation were excluded from the analysis.

To correct for the lactation stage dependent trend (lactation curve dynamics), the farm-average DMY was calculated per DIM and per parity group (1st, 2nd, 3rd and 4th or higher parity), and subtracted from each cow-individual DMY curve. The resulting residuals (**DMYc**) were standardized using a min-max standardization and do not contain a lactation-stage dependent time trend and accordingly, can be modelled with a linear model.

### Activity

Two-hourly step count values, as measured by SCR activity sensors (details sensors) were aggregated into daily values. Only the activity data for which also DMY data were available were retained for the analysis, and values outside the normal range of [180;1800] were removed. For this research, we were not interested in activity spikes caused by estrus behavior, and thus the cow-individual activity curves were filtered to correct these spikes. To this end, we first smoothed the daily activity values with a 1st order Savitsky-Golay filter in a 7-day window to obtain ACTs and subtracted the smoothed from the raw activity values to obtain activity residuals. The median average deviation of these residuals in a rolling window of 9 days (MAD9) was used to define a time-variable cow-individual threshold that accounts for both the activity level and its day-to-day variability. Specifically, activity values that exceeded the cow-individual smoothed mean activity level plus 4\*the MAD9, with a minimum activity increase of 20% were set to the ACTs+MAD9 before the analysis.

Initial data exploration showed that within farms, the average, standard deviation and range sometimes differed over time (some datasets contained data of more than 10 years). These changes were not linked with e.g. season or the number of animals, but probably resulted from hardware replacements or software updates. Additional to removing the estrus spikes, daily activity values were standardized to obtain time series with similar statistical properties over time within farm. To this end, we determined the changepoints in the median, 10th and 90th quantile using a dynamic programming search algorithm and linear cost that detects mean-shifts in these variables using least-square deviations. The activity data were then corrected with a z-score standardization for each period in which the statistical properties remained approximately constant to obtain ACTc, used in the statistical models.

### Weather information

The weather information included in CowBase is obtained from KMI in Flanders, KNMI in the Netherlands, XXX in Italy [***needs citations/links***]. For each farm, we used data from the 3 closest weather stations, as calculated by the Euclidian distance between the geological coordinates of the weather stations and the farm. Degrees longitude were taken as 111 km, whereas degrees latitude were converted to 70 km before the calculations. Daily average temperature and relative humidity values were combined according to an inverse distance weighting scheme in which the weather measured at a station further away had less impact on the final value as compared to weather measures at a station close to the farm.

Finally, the temperature and relative humidity were aggregated into a daily average Temperature-Humidity-Index (**THI**) according to the equation as published by the National Research Council (1971), and shown in eq **E1**.

(eq. E1)

With *THI* the average temperature-humidity-index for that day, the average temperature in degrees Celsius and the relative humidity, i.e. the amount of water vapor present in the air expressed in % of saturation for that temperature. Before entering THI in the models, the values were standardized *THIs* using a *z*-score standardization to obtain a similar order of magnitude as the covariates, necessary for model convergence. The formula used (eq. **E2**) was:

(eq. E2)

With *THIs* the standardized and scaled average THI per day, the average *THI* of the entire dataset for that farm and the corresponding standard deviation.

Modelling the effect of temperature on the milk yield requires the inclusion of an additional weather variable describing the lagged effect caused by heat stress. This variable, we designed based on a visual exploration of the data in which we assessed the time lag between high temperatures and the succeeding drop in milk yield. Specifically, the ‘lagged heat stress effect’ (**LHE**) was formulated as the twice the amount of hours that the temperature was above 25° and 0.5 times the amount of hours that the temperature was between 21and 25°C in the 4 days preceding the current day (eq. E3). With this variable, we assume that the effect of temperature is present already from temperatures as low as 21°C, but also that it is nonlinear, being proportionally higher with higher temperatures. Formulation of these assumptions in the LHE explicitly allows to use a model that is linear in the parameters for a variable with a nonlinear effect on the dependent variable.

LHE = (Eq. E3)

With LHE the time lagged effect of temperature, *d* the number of hours between 21°C and 25°C on day *d* and the number of hours above 25°C on day *d*. This variable was standardized with a min-max standardization before entering it in the models.

## Statistical models

To study the relation between DMY, ACT and THI, mixed linear models were developed that allow capturing both group and individual cow trends and differences. The response variables were DMYc and ACTc. Covariates were included in the design matrix to account for lactation stage (**LS;** DIM 1-21, 22-60, 61-120, 121-200 and 200+), year (**Y,** values dependent on the farm) or year-season (YS with seasons “winter” = Dec-Feb, “spring” = Mar-May, “summer” = Jun-Aug, and “autumn” = Sep-Nov) and parity effects (**PAR**; parity groups 1, 2, and 3+), which were all entered as categorical variables. We included both a random intercept and random THI-slope to allow individual variations across lactations, and modelled these with an unstructured variance-covariance matrix. The model selection to obtain the final model depended on inspection of the residuals in function of THI and time, convergence of the model and evaluation of the model fit. A Wald test was performed to assess validity of the fixed effects with a significance level of α = 0.05.

### Daily milk yield

Changes in DMYc  in function of the THIs are modelled as shown in eq. **E4**:

(eq. **E4**)

with DMYc the herd-parity corrected daily milk yield, ***1*** the intercept, LHE the variable describing the number of hot temperature hours in the previous 4 days as described before, LS, PAR and YS the covariates of respectively lactation stage, parity and year-season groups. The *id* variable stands for the cow-lactations, for which individual random effects were estimated. The random intercept and THI terms are assumed normally distributed ~ N(0,Σ1) with zero mean and an unstructured variance-covariance matrix Σ. The error term ε ~ N(0,σ1²) is the random residual error.

### Activity

Changes in ACTc  in function of THI are modelled as shown in eq. **E5**:

(eq. **E5**)

with ACTc the estrus-time corrected standardised daily activity, ***1*** the intercept, LS, PAR and Y the covariates of respectively lactation stage, parity and year groups. To model the non-linear effect of THIs on the daily activity (with more extreme THI values, proportionally higher activity increases occur), a second and third order THI term are included in the model. The *id* variable stands for the cow-lactations, for which individual random effects were estimated. Also here, the random intercept and THI terms are assumed normally distributed ~ N(0,Σ2) with zero mean and an unstructured variance-covariance matrix Σ and ε ~ N(0,σ2²) denotes the random residual error term.

### Evaluation of results

To assess the general model fit , we calculated the R², as expressed in the proportion of variance explained by the model as compared to a constant model (i.e. the mean of the data), was calculated from the sums of squares (Eq X.).

With SSE the error sums of squares, calculated as the sum of the squared residuals from the fitted model, and SSTO the total sums of squares, calculated as the sum of squared residuals from the mean. Additionally, we assessed the validity of the chosen independent model variables using a Wald test that confirms whether these are collectively significant for the model, and report the p-values. As all model variables are standardized, size and sign of the model coefficients can provide an idea of the importance and direction of each variable.

To evaluate the added value of including random effects, we calculated the marginal R² and the conditional R² as the proportion of variance explained by the fixed and fixed + random effects respectively as follows, based on the formulae proposed by Nakagawa et al. (2017) (eq. EX and EX):

(eq. EX)

(eq. EX)

With the residual variance of a model with only the fixed effects included, and the variance of the random intercept and THI effects, respectively, as derived from the variance-covariance matrices Σ1 and Σ2, and the residual error of the mixed model. The correlation between the individual random effects of the intercept and slope are calculated as

With the covariance of the random intercept and THI-slope, and the standard deviations of the random intercepts and THI slopes, respectively.

# Results

C:\Users\u0084712\OneDrive - KU Leuven\projects\ugent\heatstress\dataanalysis\results\activity\sexample_ind_activity_farm_5_cow_7528_withmilk.tifFigure 1 shows an example of the milk yield and activity data of a cow in the sixt parity from farm 5, and the corresponding herd dynamics with the THI indicated in red. In the upper panel of this figure it becomes clear that the average herd activity pattern follows a similar trend as the THI, and that peaks in THI coincide with peaks in activity (light blue shadow in upper panel). Also, small dips in the milk yield (lower panel) a few days after the activity peaks appear, suggesting the time-lagged effect of THI on the milk yield. These observations are further confirmed with the model results as discussed below.

**Figure 1**. Example of herd and individual trends of the activity and milk production data. The upper panel shows an individual time series of standardized activity in green with in blue the corresponding individual trend line. The blue shadow displays the herd trend (average ± std). The red line is the average THI, whereas the pink zone indicates when THI is above 68. The lower panel shows the individual milk yield curve (purple line) and the milk yield dynamics of the corresponding parity group in the same period (light purple shadow, average ± std).

* 1. Milk production

***Table T2****. Model fit, random effect correlation, error variance and significance of the model variables as tested with a Wald Chi-squared test.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Daily milk yield** | **farm 1** | **farm 2** | **farm 3** | **farm 4** | **farm 5** | **farm 6** |
| **R²**a | 0.466 | 0.439 | 0.429 | 0.467 | 0.456 | 0.517 |
| **R²(m)**b | 0.120 | 0.173 | 0.225 | 0.141 | 0.121 | 0.178 |
| **R²(c)**c | 0.957 | 0.928 | 0.910 | 0.949 | 0.956 | 0.946 |
| **Wald test *P*-values** |  |  |  |  |  |  |
| **Intercept** | 0.053 | *<0.001* | 0.214 | 0.993 | *<0.001* | *<0.001* |
| **LS** | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* |
| **Year** | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* |
| **Season** | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* |
| **Year:Season** | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* |
| **THIs** | 0.188 | *<0.001* | 0.077 | 0.086 | *<0.001* | 0.896 |
| **THIs:LS** | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* |
| **LHE** | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* |
| **Random effects correlation** | -0.857 | -0.74 | -0.636 | -0.845 | -0.857 | -0.680 |
| **Residual error variance (ε)** | 0.343 | 0.398 | 0.387 | 0.356 | 0.354 | 0.284 |
|  |  |  |  |  |  |  |

a coefficient of determination, calculated as 1-SSE/SSTO; b marginal R-squared indicating the variability explained by a linear model with fixed effects only; c conditional R-squared indicating the variability explained by the mixed model with both fixed and random effects.

The mixed model with the THI, the time-lagged effect of temperature LHE, the covariates and the cow-individual random effects explains 43 to 52% of the variation in the data. The THI only affects daily milk production significantly on the same day in 2 farms, whereas for all farms, the effect of THI differs across lactation stages. The effect of parity was tested as well, but shown not to be significant for all farms. Furthermore, year, season and year\*season all significantly affect the daily milk production of the herd. By including a time-lagged effect of heat stress, that was strongly significant for all farms, we demonstrate that the effect of warm temperatures are often time lagged, as also seen by visually inspecting the individual cows’ time series data (see also Figure 1). As shown by the difference between the marginal and conditional R², the milk yield level and THI slope of individual cows varies significantly. More specifically, the random effects correlation varies between -0.857 and -0.636, indicating that cows with higher milk yield as compared to their herd mates in similar years, seasons and lactation stages, suffer more from a high THI. More specifically, when a cow produces more milk, its decline in milk production with increasing THI is larger than for a cow with a lower daily milk production.

* 1. Activity

Table T3

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Daily activity** | **farm 1** | **farm 2** | **farm 3** | **farm 4** | **farm 5** | **farm 6** |
| **R²**a | 0.511 | 0.528 | 0.446 | 0.542 | 0.490 | 0.616 |
| **R²(m)**b | 0.490 | 0.483 | 0.480 | 0.484 | 0.452 | 0.474 |
| **R²(c)**c | 0.827 | 0.845 | 0.787 | 0.841 | 0.824 | 0.883 |
| **Wald test *P*-values** |  |  |  |  |  |  |
| **Intercept** | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* |
| **LS** | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* |
| **PAR** | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* |
| **Year** | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* |
| **Season** | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* |
| **Year:Season** | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* |
| **THIs** | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* |
| **THIs:LS** | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* |
| **THIs²** | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* | *<0.001* |
| **THIs³** | 0.009 | 0.008 | *<0.001* | *<0.001* | 0.204 | *<0.001* |
| **Random effects correlation** | 0.158 | 0.130 | 0.082 | 0.156 | 0.087 | 0.325 |
| **Residual error variance (ε)** | 1.722 | 3.137 | 1.926 | 2.688 | 4.236 | 0.679 |

a coefficient of determination, calculated as 1-SSE/SSTO; b marginal R-squared indicating the variability explained by a linear model with fixed effects only; c conditional R-squared indicating the variability explained by the mixed model with both fixed and random effects.

Between 44 and 61% of the variation in daily activity is explained by the THI variables and model covariates. Wald test statistics suggest that daily activity differs across lactation stages, parities, years and seasons for all farms. THI affects the daily activity as well. The first and second order THI parameter estimates had a positive sign for all farms, suggesting that activity increases with increasing THI. Significance of the second and third order terms suggest that changes in activity are nonlinear in the parameters. The effect of THI on activity differs across lactation stages.

Also for activity, the difference between the marginal and conditional R² suggests the large variability in daily activity across individual animals. Yet here, the random effects correlation is positive and varies between 0.08 and 0.32, indicating that cows that are in general more active than average increase their activity (slightly) more as compared to cows with a lower baseline activity level. The large difference in residual error between the farms mainly demonstrates that the obtained time series differ significantly across farms despite the preprocessing measures. This can result from difference in data quality and quantity, (dis)continuity of the sensors, management differences or other (unknown) factors.

# Discussion

## Milk production

### Covariates

### Effect of THI

### Individual cow correlations

## Activity

### Covariates

### Effect of THI

### Individual cow correlations

## CowBase

* Role / added value: time won by not having to clean, simply applying criteria, easy visualization of similar traits
* Merit of this study: large datasets, longitudinal data, multiple farms, multiple countries
* Discussion of results across farms and countries

## Future perspectives

* Applications and research applied to heatstress and THI / combining act and dmy / different weather features /
* Development, use, distribution of CowBase

# References

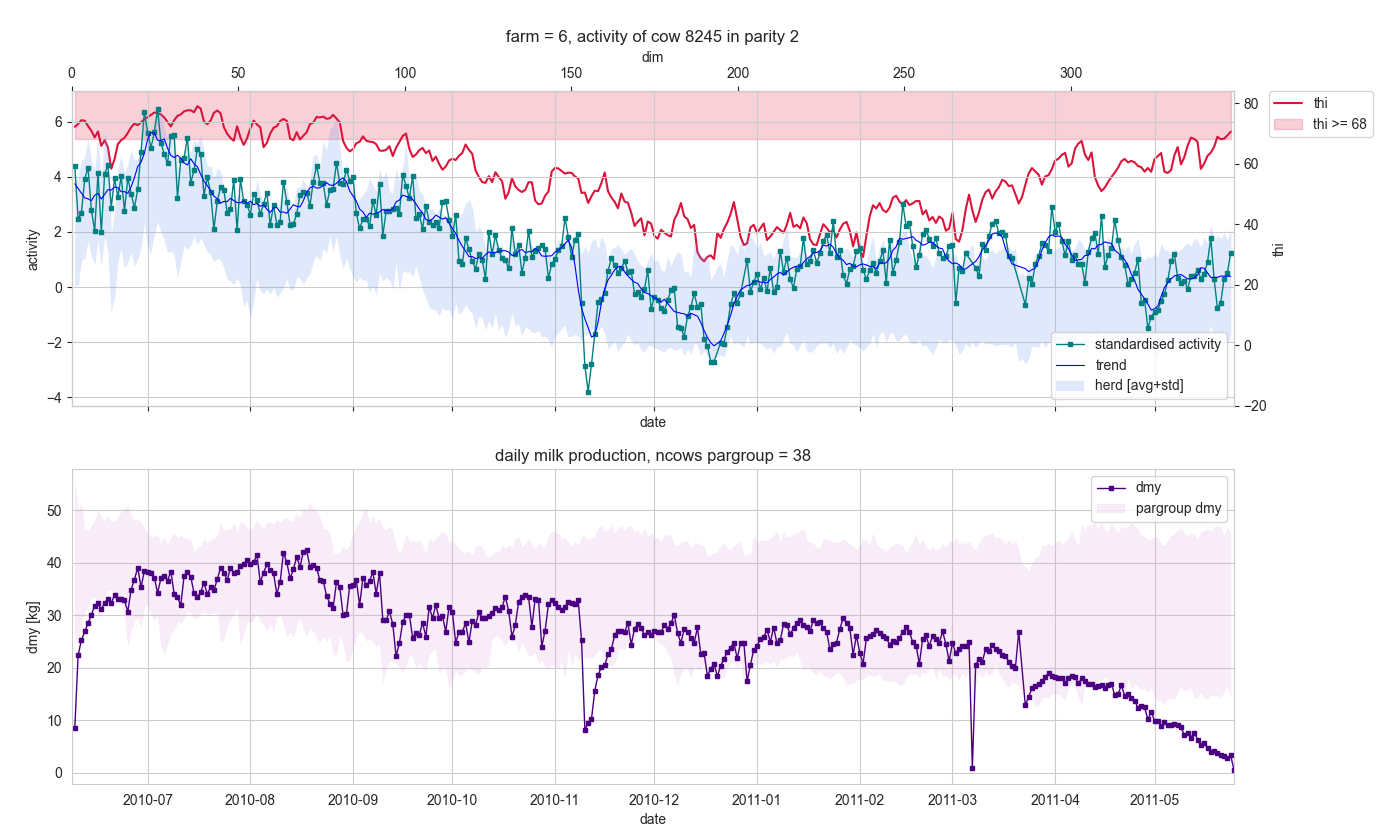
National Research Council, 1971. “A guide to environmental research on animals. National academies.

[mentioned in <https://www.sciencedirect.com/science/article/pii/S0022030207716818>]

**notes**

https://www.ssc.wisc.edu/sscc/pubs/MM/MM\_TestEffects.html

C:\Users\u0084712\OneDrive - KU Leuven\projects\ugent\heatstress\dataanalysis\results\activity\activity_farm6_example_DMY_ACT_drop2.tifnot very scientific, but this is what we can/could be analyzing: the concurrence of drops in ACT and DMY, eg end of March 2011 for this animal, and beginning of june 2011



Motivation to leave farm 5 out for now:

* Left figure is the average activity corrected, daily means. Seems OK
* Right figure is the corresponding standard deviation. In principle, the correction should have rendered a similar mean + std throughout the dataset, but for an unknown reason I have now no time to figure out, it didn’t work and the std shows huge jumps for this farm. Can be as simple as I indicated wrong number of changepoints or so, but this is not as expected
* I suggest to, for now, leave this farm out for the case and I will check later what went wrong here.

https://medium.com/@analyttica/understanding-wald-test-2e3fa7723516#:~:text=%CE%B20%20%3A%20Parameter%20of%20interest%2C%20usually,different%20than%20zero%20or%20not.&text=The%20Wald%20test%20results%20interpretation,and%20the%20variable%20is%20significant.