**Classification of environmental factors potentially motivating for dairy cows to access shade**

Matheus Deniz[1,2], Karolini Tenffen de Sousa[1,2], Isabelle Cordova Gomes[2], Marcos Martinez do Vale[2], João Ricardo Dittrich[1,2]

[1]Programa de Pós-Graduação em Zootecnia, Departamento de Zootecnia, Universidade Federal do Paraná, Curitiba, Brazil.

[2]Laboratório de Inovações Tecnológicas em Zootecnia (LITEZ - UFPR), Departamento de Zootecnia, Universidade Federal do Paraná, Curitiba, Brazil.

**SUPPLEMENTARY FILE**

## Materials and Methods

*Experimental area and climate pattern*

This work was carried out in a silvopastoral system (SPS) on a commercial dairy farm in southern Brazil. Data collection was performed during summer (southern hemisphere); in four consecutive days with high temperatures, high solar radiation, and low cloudiness. According to Köppen classification, the climate of the region is subtropical humid mesothermic (Cfa) and presents hot summers with average annual temperatures between 18 and 20°C and relative humidity between 63 and 84% (INMET *et al*., 2009; Alvares *et al*., 2013).

The experimental area had 4 paddocks (1.550m²/ paddock) where each one was composed of a silvopastoral system. This system consisted of native trees (approximately 8 meters high) planted in wood with a distance of 14 meters, and provided a total shaded area of 5m²/ animal in each paddock (determined by Shading Vegetation Index) and a sunny area of 33m²/ animal in each paddock. At the farm, animals are raised permanently on pasture, mainly composed of plant species of *Axonopus catarinenses, Arachispintoi spp.* and *Paspalum notatum*. The pasture is managed under Voisin's Rotational Grazing system whereby animals are moved daily to a new paddock. Thus, as the paddocks and SPS distribution were uniforms, this allowed us to evaluate one paddock per day.

*Animals and frequency at the shaded and sunny areas*

Lactating Jersey cows (n = 39), with similar coat colour (light brown), weight (mean ± SD) of 450 ± 50kg were observed during four days, for 8h each day (from 09:00 to 16:50). All observations were performed in an area already known by the animals and began after the last animal entered at the paddock. To minimized research bias, after milking morning, animals were handled by farmers to the experimental area. Frequency of animals in each area (shaded and sunny) was recorded by scan sampling of 10 min. intervals (Altmann, 1974). The cow was considered to be in the shaded area when more than 50% of her body was in the shade of the tree. The cow was considered to be in the sunny area when more than 50% of her body was in the sun (Kendall *et al*., 2006; *Giro et al*., 2019). All observations were made by researchers previously trained and with knowledge in the area of animal behaviour; in order to not interfere with the animals' behaviour, the observations were performed outside of the paddock with a safe distance. The reliability of simultaneous observations of a given individual by the observers reached 94.2% before the beginning of the data collection.

*Environment evaluation*

During the experimental period, environmental factors were collected in 120 points [fifteen in each area (shaded and sunny)]. Thus, in order to avoid temporal variations between the areas, data collection was carried out simultaneously in both areas. In shaded and sunny areas of the SPS, the following environmental factors were measured: air temperature (AT, °C), relative humidity (RH, %), solar radiation (SR, W/m²) and wind speed (WS, m/s).

Air temperature (°C) and relative humidity (%) measurements were performed (with solar radiation shield) with a thermo-hygrometer (humidity 0-100% scale; ± 2.5% accuracy; 0.1% resolution; temperature, -30 to 100°C scale; ± 0.8°C accuracy; and 0.1°C resolution). The solar radiation measurement was performed with a pyranometer (0 to 4000W/m²; ± 4% accuracy). Wind speed was measured with a thermo-anemometer (0.4 to 20 m/s scale; ± 2% accuracy). Data collection was carried out from 9:00 to 16:50 at a height of 1.3m from the ground (height average of the center of mass of Jersey adult cattle) with intervals of 10 min., and averages were generated every 1 h.

*Data mining and statistical analysis*

Animal frequency at the areas and environmental data were used to build a database with 29320 observations and 10 variables, one being the classification (Table S1). The database was built with each observation (frequency at the areas and environmental) synchronized by date and time of day. Data mining technique was applied following CRISP-DM methodology (Klein *et al*., 2020).

**Table S1**

Data mining was performed with the software Waikato Environment for Knowledge Analysis (WEKA®, 3-4), which classifies the data and build a classification tree using the J48 algorithm, an implementation of the algorithm C4.5 that is a supervised machine learning tool. The J48 algorithm generates a model with semantic rules using the minimum information required for classification. The model result is expressed graphically in the form of an inverted tree; the first attribute is the one with the highest classification power (root node). From the root node, semantic rules are expressed as body ➔ head. The rules body are logic connectors (≤, >, and =) called as nodes that express the connection between the features that are capable to classify an event. The classification from a rule is the head that is represented in the graphic tree as the leafs. Each branch in the classification tree is one rule with their connectors in the body and a class on the head.

Classification tree was generated by ranking the cow's frequency at the areas (shaded or sunny), according to the environmental factors. The best model selection was based on the model accuracy, the precision of classes, and the interpretation of classification rules by experts with the minimum requirement of three years of expertise. In the analysis, were applied a ten-fold cross-validation, available in the J48 algorithm. Model accuracy, as well as class precision, were calculated by a confusion matrix (Table S2). The class precision ranges from zero to one and expresses the relation of true positive and true negative classifications in a specific class.  The model accuracy expresses the percentage of instances that were correctly classified.

**Table S2**

In order to confirm the level of agreement of the data sets and classification accuracy, the Kappa statistical method was used (see more information in: Sim and Wright 2005; McHugh 2012) was determined by equation (1) developed by Cohen (1960). In this study, when describing the relative strength of agreement associated with kappa statistics, the labels proposed by Landis and Koch (1977) were used. The relative strength values indicate: ≤0: poor; 0.00 – 0.20: slight; 0.21 – 0.40: fair; 0.41 – 0.60: moderate; 0.61 – 0.80: substantial; and 0.81 – 1.00: almost perfect.

$$\mathrm{K} = \frac{\mathrm{P_0} - \mathrm{P_c}}{1 - \mathrm{P_c}} \qquad (1)$$

Where:

K is the kappa statistical,

$P_o$ is the proportion of observed agreements and,

$P_c$ is the proportion of agreements expected by chance.

As confirmatory analysis, the data (frequency at the areas and environment) were submitted to the normality test (Shapiro-Wilk), analyzed by Generalized Linear Models (GLM) and submitted to the Spearman correlation test. Experimental design of environmental factors was composed of four replicates (paddocks), 120 experimental units (30 collection points by paddock), two independent variables (shade and sun) and four dependent variables (air temperature, relative humidity, solar radiation and wind speed) following the model:

$$Y_{ij} = \alpha_j + \beta_{ij} + e_{ij}$$

Were:

$Y_{ij}$ are the microclimatic variables,

$\alpha_j$ are the fixed effect of the areas provided by the silvopastoral system,

$\beta_{ij}$ is the random effect, *i* corresponds to days; *j* corresponds to hours, and

$e_{ij}$ is the residual effect.

All analyzes were performed separately and each environmental factor obtained a GLM model. Gamma distribution and logarithmic bonding function were used for the environmental factors, at a 95% confidence level.

The analysis of frequency at the areas was composed of four repetitions (paddocks), 39 experimental units (animals), two independent variables (shade and sun) and the dependent variable was the frequency of events recorded in shaded and sunny areas. Poisson distribution at a confidence interval of 99% was used. Animals, days and hours were defined as random effects following the model:

$$Y_{ij} = \alpha_j + A_i + \beta_{ij} + e_{ij}$$

Were:

$Y_{ij}$is the cow's frequency at the areas,

$\alpha_j$ are the fixed effect of the areas provided by the silvopastoral system,

$A_i$ is the random effect of animals,

$\beta_{ij}$ is the random effect, *i* corresponds to days; *j* corresponds to hours, and

$e_{ij}$ is the residual effect.

All analyzes were performed through the statistical software R (R Core Team 2019) and all statistical models were adjusted using the maximum likelihood-Laplace approximation method in the statistical package lme4 (Bates *et al*., 2015).

## References


**Altmann J** (1974) Observational Study of Behavior : Sampling. *Behaviour* **49**, 227–267. doi:10.1080/14794802.2011.585831.

**Alvares CA, Stape JL, Sentelhas PC, Gonçalves JLM, Sparovek G** (2013) Köppen's climate classification map for Brazil. *Meteorologische Zeitschrift* **22**, 711–728. doi:10.1127/0941-2948/2013/0507.

**Bates D, Mächler M, Bolker B, Walker S** (2015) Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software* **67**, 1–47. doi:10.18637/jss.v067.i01.

**Cohen J** (1960) A coefficient of agreement for nominal scales. *Educational and Psychological measurement* **46**, 20–37.

**Giro A, Pezzopane JRM, Barioni Junior W, Pedroso AF, Lemes AP, Botta D, Romanello N, Barreto AN, Garcia AR** (2019) Behavior and body surface temperature of beef cattle in integrated crop-livestock systems with or without tree shading. *Science of the Total Environment* **684**, 587–596. doi:10.1016/j.scitotenv.2019.05.377.

**INMET** (2009) 'Normais Climatológicas do Brasil.' (INMET: Brasília)

**Kendall PE, Nielsen PP, Webster JR, Verkerk GA, Littlejohn RP, Matthews LR** (2006) The effects of providing shade to lactating dairy cows in a temperate climate. *Livestock Science* **103**, 148–157. doi:10.1016/j.livsci.2006.02.004.

**Klein DR, Vale MM, Silva MFR, Kuhn MF, Branco T, Santos MP** (2020) Data mining as a hatchery process evaluation tool. *Scientia Agricola* **77**, 2018–2021. doi:10.1590/1678-992x-2018-0074.

**Landis JR, Koch GG** (1977) The Measurement of Observer Agreement for Categorical Data.

*Biometrics* **33**, 159–174. doi:10.2307/2529310.

**McHugh ML** (2012) Lessons in biostatistics interrater reliability : the kappa statistic. *Biochemica Medica* **22**, 276–282. https://hrcak.srce.hr/89395.

**R Core Team** (2019) R: A language and environment for statistical computing. R Foundation for Statistical Computing Vienna, Austria.

**Sim J, Wright CC** (2005) The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy* **85**, 257–268. doi:10.1093/ptj/85.3.257.

**Supplementary tables legends:**

**Supplementary table 1:**

Summary of data and variables of the final database.

**Supplementary table 2:**

Confusion matrix representation.

**Supplementary table 1**

| Nº | Variable | Unit | Nº | Variable | Unit |
|---|---|---|---|---|---|
| 1 | Day[A] | Numeric | 6 | Air temperature | °C |
| 2 | Hour[B] | Numeric | 7 | Relative humidity | % |
| 3 | Categorized time[C] | Numeric | 8 | Solar radiation | $W/m^2$ |
| 4 | Scan[D] | Numeric | 9 | Wind speed | m/s |
| 5 | Animals ID[E] | Numeric | 10 | Areas: shaded/ sunny[F] | Class |

[A]collection days; [B]hours of data collection (range: 1 to 8); [C]categorization of observation hours in period (morning and afternoon); [D]observations of frequency at the areas in each 10min.; [E]individual identification by animal; [F]nominal classification of each event based on the area used by animal.

**Supplementary table 2**

| Class | Predict as $C_+$ | Predict as $C_-$ | Class precision | Model accuracy[A] |
|---|---|---|---|---|
| $C_+$ | True positives ($T_p$) | False negatives ($F_n$) | $T_p/ (T_p + F_n)$ | $[T_p + T_n)/ N] \times 100$ |
| $C_-$ | False positives ($F_p$) | True negatives ($T_n$) | $T_n/ (F_p + T_n)$ | |

[A]N is equal to the number of instances in the test set.