null • Issue 1

Predicting Varroa: Longitudinal Data, Micro Climate, and Proximity Closeness Useful for Predicting Varroa Infestations (I1.A1)

**Michael Rubinigg¹, Megan MacDonald², Veronika Davenport²,
Ed Hassler², Awad Hassan³, Vitore Shala-Mayrhofer⁴, Joseph Cazier²**

¹Center for Analytics Research & Education, Appalachian State University, USA ; Biene Österreich (Austrian Beekeeping Federation), Austria,

²Center for Analytics Research & Education, Appalachian State University, USA,

³Center for Analytics Research & Education, Appalachian State University, USA ; Department of Plant Protection, South Valley University, Egypt,

⁴Austrian Chamber of Agriculture (LKÖ), Vienna, Austria

Published on: Feb 13, 2021

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

Structured Abstract

Data Overview: An eight-year survey of Varroa destructor infestation rates of western honey bee (*Apis mellifera*) colonies across Austria and the spatial dimension, temporal dimension and weather factors that impact these infestation rates.

Data Value: This data is valuable because it was collected and designed to build predictive models on where Varroa is likely to appear based on weather, geography, and other factors; and enhances the varroa population models. Bees are important because humans rely on honey bees to pollinate many of the earth's food crops. *Varroa destructor* is responsible for 30% of western honey bee deaths. No honey bees, no pollination of food crops. Aligning with the UN sustainability goals *15: Life on Land* and secondly *2: End Hunger*, this dataset provides tools to predict colony loss of the western honey bee population; allowing beekeepers to implement disaster risk reduction strategies to reduce biodiversity loss and ensure pollination of the world's food crops.

Data Description: The data contains 6 tables including user, yard, hive, varroa sampling, station and weather data. Varroa samples were collected from citizen and mined data using 3 standard method sources between the years 2012-2020. There are 99 unique user_id's, 242 yard_id's, 2,116 hive id's, totalling to 11,124 unique varroa sampling events. The combined hourly weather data collected from 73 weather stations around Austria for the 8 years, total ~1.3 million rows.

Data Application: This data is suited for predictive modeling, decision trees, neural networks and linear regressions as well as both unsupervised and supervised learning projects.

Possible research questions include but not limited to:

- What impact does seasonality have on the varroa infestation of western honey bees in a given environment?
- How have weather factors affected infestation rates?
- What is the relationship between proximity to other hives and infestation rates?
- What are the main spatial and temporal dimension factors that influence Varroa infestations?

Indexing Table

Supported UN SDGs	15: Life on Land, 2: End hunger
Type of Data/Article	Archival Data - Not continuously updated
Class of Analytics	Predictive
Number of and List tables	Six tables: <i>user</i> , <i>yard</i> , <i>hive</i> , <i>varroa_sampling</i> , <i>station</i> , and <i>weather</i> .
Key words	Varroa, beekeeping, ectoparasite, ecological modeling, infestation modeling

Introduction

Varroa destructor is an ectoparasitic mite, and currently poses one of the most serious threats to the western honey bee (*Apis mellifera*). This mite has invaded western honey bee colonies throughout the world, except in Australia. The dataset in this article consists of an eight-year survey of varroa infestation rates of honey bee colonies in Austria and the potential factors that impact these infestation rates. The dataset has been built on citizen and mined data. The varroa survey data was collected by software and a mix of trained and untrained beekeepers. Data on factors most likely to impact varroa infestation levels are included in the dataset. These factors are in two categories: 1) beekeeping factors which include the region and elevation of bee yards (the spatial dimension), and time of the year (temporal dimension) as well as 2) weather factors which include hourly values of air temperature, dew point, pressure, wind direction, wind speed, sky conditions, and precipitation. The dataset has been processed through professional data management steps including filing, cleaning, transforming, and anonymizing. This dataset may be useful (solely or with other datasets) for researchers and decision-makers to build models to predict the impact of climate change on varroa, enhance the varroa population models, and understand the population dynamics of varroa and ectoparasites in general.

U.N Goal ID

Primary: *15 life on Land: Protect, restore, and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, and halt and reverse land degradation and halt biodiversity loss.* [7]

Secondary: *2 End hunger: Achieve food security and improved nutrition and promote sustainable agriculture.* [8]

The dataset provides a tool in aiding disaster risk reduction strategies and climate change adaptations when implemented; leading to the reduction of biodiversity loss of the western honey bee population. As reaffirmed in the Rio+20 conference “the value of biological diversity plays a critical role in maintaining ecosystems that provide essential services, which are critical foundations for sustainable development and human wellbeing”[9]. Honey bees play a crucial role in pollinating many of the earth's food crops such as fruits and nuts; a loss in western honey bees would decrease pollination and increase food insecurity in those affected regions. The data provided will aid beekeepers in implementing preventative control measures to protect western honey bees, which promotes the sustainable use of terrestrial ecosystems and agriculture and aligns with U.N Sustainability Goal 15 and 2.

Value of Data

This dataset was collected, designed and expressly intended to build a predictive model to predict in advance where varroa was likely to appear based on weather, geography, and other factors and is ideally suited for that purpose. Additionally, given that this data was collected over 8 years with regular updates from each apiary during each year this data is valuable to explore long-term varroa cycles both within a year and across several years within apiaries. With multi-year and multi-month data collection combined with hourly weather data collected within a few kilometers of the hives, this data is ideally suited to illuminate the impact of weather on varroa infestations. Finally, given this data was collected in a mountainous region with hives at many different elevations with different flora, fauna, and seasonality, it provides a rich set of control factors for studying the attributes listed above while also exploring the role of elevation on varroa infestations.

Data Summary

Data for this project are in .csv files. Files in this dataset are: *user*, *yard*, *hive*, *varroa_sample*, *station*, and *weather*. The *yard_id* of file *varroa_sampling* joins with the file *yard*. The *hive_id* of file *varroa_sampling* joins the file *hive*. The *user_id* of file *hive* joins the file *user*.

Weather data is derived from NOAA and stored in the *weather* file. It can be joined to the *yard* file through the *station* file using the *station_id*, which matches a bee yard to the closest weather station. The *station* file is joined to the *weather* file by the *station_id*.

More information about the attributes provided in all of the files, their formats, and a brief description are contained in the **Entity Data Dictionary [Last Section : Entity 1-6]**. **Figure 1** shows the joins among the 6 files of this dataset. **Figure 2** shows mapped locations of weather stations included in this dataset.

For reference, the format column contains the following abbreviations for data formats

- *varchar* - a variable-length string of text characters
- *int* - numeric integer values
- *double* - numeric double-precision floating-point values
- *datetime* - values represent the date and time in the local timezone
- *string*- sequence of characters
- *serial*- unique identifier assigned incrementally or sequentially to an item

Entity Relationship Diagram

Figure 1: Entity Relationship Diagram of tables. Entities are joined to those with an arrow between them based on common columns.

Figure 2: A map showing the weather stations' locations and the number of bee yards associated with each station (circles). Regions belonging to the same NUTS-1 level (AT1: yellow; AT2: red; AT3: green) have the same colour.

Enhanced Analysis Data

Data were explored using descriptive statistics. **Table 1**, below further describes some of these metrics for selected key columns. Overall, there are 2,116 unique hives in this dataset contained in 242 yards and managed by 99 individual beekeepers.

The bee yards are between an elevation of 150 meters to 1,413 meters above sea level. These yards are all in Austria, and the majority of them are on the eastern side of the country.

Table 2 describes weather data temperature, dew point, pressure, wind direction, wind speed, sky conditions, and precipitation. To note in weather data: variables of -9999 indicates the missing value, -1 in precipitation indicates trace precipitation. Temperature ranges between -16 and 34.7 degrees celsius. Dew point ranges from -31 to 27.5 degrees celsius. Pressure ranges from 999.2 to 1042.8. Wind speed ranges from 0 to 28 m/s. Precipitation for 1 hour ranges from 0 to 69 measured in **.1 centimeters**. Precipitation for 6 hour ranges from 0 to 99 measured in **.1 centimeters**. Sky conditions range from 0 to 9. Please see **Chart 1** for the frequency of Sky Conditions. Oktas 8 which represents full cloud cover with no breaks occurs most frequently while 9 oktas which represents a sky obscured by fog or other meteorological phenomena occurs least frequently.

Yards connect to the weather files by the closest weather station. There are 73 unique weather stations in this dataset. These are contained between latitudes of 46.617 and 48.683 and longitudes of 9.617 and 16.600. Because of the public availability of this data, these coordinates are not blurred. These stations can be found between elevations of 153 meters and 1210 meters above sea level. 90% of the yard elevations are within 300 meters of the weather station elevation. This means there is between a 0 and 6 degrees celsius difference in air temperature which can be calculated with the data provided for accuracy in analysis.

This data contains 11,124 varroa sampling events. Further information on how the varroa samples are collected can be found below. Roughly 21% of these events record zero mites present. The highest number of mites present in a single sampling event is 5,016 Sampling events are collected from 04/02/12 to 11/11/20, and last on average 7.2 days each (range: 1,0-23.8 days). **Chart 2** represents the frequency of day ranges for sampling events. ~75% of Varroa Sampling events occur between 3 and 9 days.

Table 1: Frequency of Key Data Columns

Attribute Name	Frequency
<i>sampling_id</i>	11,124
<i>hive_id</i>	2,116
<i>yard_id</i>	242
<i>user_id</i>	99
<i>station_id</i>	73

Table 2: Descriptive Statistics of Weather Data (missing values, [-9999] and trace precipitations [-1] were excluded to run summary statistics, however, included in original datasets)

Attribute Name	Mean	Standard Deviation	Standard Error	Min	Max
<i>Air Temp (Celcius)</i>	10.15	8.72	0.01	-23.00	36.60

<i>dew point</i>	5.28	7.24	0.01	-31.00	27.50
<i>precip 1 hr</i>	0.75	1.67	1.00	0.00	69.00
<i>precip 6 hr</i>	1.05	3.26	0.01	0.00	99.00
<i>pressure</i>	1017.71	8.35	0.01	943.9	1050.9
<i>sky conditions</i>	n/a	n/a	n/a	0.00	9.00
<i>wind direction</i>	194.92	116.45	0.10	0.00	360
<i>wind speed</i>	2.24	1.91	n/a	0.00	28.00

Chart 1: Frequency of Sky Condition.

Oktas are a measurement of the total celestial dome covered by clouds or other obscuring phenomena.

0 oktas represent the complete absence of cloud,

1 okta represent a cloud amount of 1 eighth or less, but not zero,

7 oktas represent a cloud amount of 7 eighths or more, but not full cloud cover,

8 oktas represent full cloud cover with no breaks,

9 oktas represent a sky obscured by fog or other meteorological phenomena.

Cloud covering chart for reference:

<https://polarpedia.eu/en/okta-scale/>

Frequency of Sky Condition

Chart 2: Frequency of Day Range of Varroa Sampling Event. The length in days of each Varroa sampling event occurred. Majority of samples occur between 3 - 9 days with the other 25% ranging between 1-3 days and 10-24 days.

Frequency of Day Range of Varroa Sampling Event

Varroa Survey Data Methods

Data on mite infestation levels were collected from 3 sources by a standard method - natural mite falls [5] - from 2012 to 2020, mainly in the spring, early summer, and late

summer. Data were collected from 3 different sources of differing quality. Data from the highest quality, described as `quality_control=2`, was examined with the BeeVS diagnostic system (Apisfero, Turin, Italy), which consists of a high-resolution scanner to take a picture of the samples (sticky boards placed under the brood nest of colonies) and cloud-based software used to count the number of mites on the sticky boards. Data from the intermediate source is described as `quality_control=1` and were examined manually by a trained group. Data from the poorest quality source is described as `quality_control=0` and were examined manually by untrained individuals according to a classification scheme. Data was entered via a web terminal by whomever analyzed the sample. The software vetted the data for plausibility (rejection of values that exceed 100 mites/day) and completeness (rejection of values that did not fall between a 3 day and 21-day measuring interval). Data exceeding these limits, which can be found in the data set, has been imported from external resources and has been approved by the supervisor. The data collected by untrained individuals were checked by the supervisor for plausibility.

From 2012 to 2016 the project was only implemented in the Austrian province of Styria, where approximately 3,500 beekeepers supervised 53,000 to 56,000 honeybee colonies. In 2017 the crowdsourcing initiative was extended to all nine Austrian provinces, consisting of 28,032 to 30,237 beekeepers and 329,402 to 390,607 honeybee colonies in their care [6].

The total number of samples collected is 11,124. 4,033 (36%) were medium quality samples (`QC=1`) and 3,267 (29%) were high quality samples (`QC=2`). **Chart 3** It is important to note that there is a strong bias in the origin of the samples. A single user provided 27% of the samples in this dataset. About 53% of the samples are derived from 22 users who each provided 100 to 999 samples. 18% of the samples are from a group of 56 users who provided 10 to 99 samples. 1% of the samples were given by 24 users who had entered less than 10 samples each.

The varroa survey dataset includes 99 users (beekeepers), 242 bee yards (apiaries), and 2,116 hives from the nine Austrian provinces for a total of 11,124 records pertaining to varroa infestation.

Chart 3: Frequency of quality control in Varroa sampling Chart. This chart shows the frequency of each quality control.

Frequency of Quality Control

Weather Data Methods

Weather data is derived from NOAA, using Integrated Surface Data Lite (ISD-Lite). The ISD-Lite data contains a formatted subset of the complete Integrated Surface Data (ISD) for a number of elements. The data are based on data exchanged under the World Meteorological Organization (WMO) World Weather Watch Program according to WMO Resolution 40 (Cg-XII). The data of the Austria weather stations have been filtered from: <ftp://ftp.ncei.noaa.gov/pub/data/noaa/> by unique USAF, WBAN, and year. The hourly values of temperature, dew point, wind speed, pressure, and precipitation have been maintained in the data set and preserved in original metric measurements. Each bee yard has been matched to the closest weather station. The dataset includes 73 weather stations, 2012-2020 hourly values, and 1.3 million records. [3]

Data Management

Data has been separated into the above tables and organized based on matching keys. Data that was too revealing of specific beekeeper identity or locations (zip code, lau code, coordinates) was removed from the table yard for bee yards. This was done to

protect beekeepers in this study. Weather data was cleaned to only contain observations that fall between 2012 and 2020, and only for the stations closest to bee yards included in the Varroa survey. Bee yards were joined to nearby weather stations using various packages in R software. All variables were reformatted to be measured in standard metric units wherever they were not.

Data Application and Conclusion

Varroa destructor, the ectoparasitic honey bee mite, is a driver for colony losses throughout the world, except in Australia. With this data, users can implement unsupervised and supervised learning to further this discussion of Varroa's threat to the western honey bees. This dataset aims to provide tools in building models to predict the impact of climate change on varroa, to enhance the varroa population models, and to understand the population dynamics of varroa and ectoparasites in general.

Possible Questions and projects of explorations within the dataset:

- What impact does seasonality have on the varroa infestation of western honey bees in a given environment?
- How have air temperature, precipitation, wind speed, wind direction, air pressure, and sky conditions affected infestation rates of hives across Austria in a given year?
- What is the relationship between proximity to other hives and infestation rates?
- What effect has climate change had on varroa cycles in Austria?
- What are the main spatial and temporal dimension factors that influence Varroa infestations?
- Map: Disease vectors for how infestations spread within yards for similar environments
- Model: Seasonal growth rates of Varroa mite populations based on regions
- Model: Prediction models of where varroa was likely to appear based on spatial, temporal, and meteorological dimensions.

With a robust set of control factors for studying the variables listed above the value of the dataset provides a wide range of data for exploration of multiple variable impacts on varroa infestations. This dataset aligns with the UN sustainability goals 15 and secondly 2 by aiding in providing tools to predict colony loss of the western honey bee population; resulting in one step closer to implementing disaster risk reduction strategies to reduce biodiversity loss and ensure pollination of many of the world's food crops.

Entity Data Dictionary

Entity 1: Description of the attributes in the *user* file

Attribute Name	Format	Description
<i>user_id</i>	varchar(4)	The user identifier
<i>samples</i>	integer	Total numbers of samples provided by a user

Entity 2: Description of the attributes in the *yard* file

Attribute Name	Format	Description
<i>yard_id</i>	varchar(4)	The yard identifier
<i>elevation</i>	decimal(5,1)	Meters above Sea Level rounded to the nearest meter
<i>nuts</i>	varchar(5)	NUTS is a geocode standard for referencing the administrative divisions of countries for statistical purposes. AT1 - East Austria; Burgenland (AT11), Lower Austria (AT12), Vienna (AT13) AT2- South Austria; Carinthia (AT21), Styria (AT22) AT3 West Austria; Upper Austria(AT31), Salzburg(AT32), Tyrol(AT30), Vorarlberg (AT34) The current Nomenclature of Territorial Units for Statistics (NUTS) adopted by the European Union (Commission Delegated Regulation 2019/1755) is applied.
<i>station_id</i>	varchar(6)	The NOAA weather station identifier

Entity 3: Description of the attributes in the *hive* file

Attribute Name	Format	Description
<i>hive_id</i>	varchar(4)	The hive identifier
<i>user_id</i>	varchar(4)	The user identifier

Entity 4: Description of the attributes in the *varroa_sampling* file

Attribute Name	Format	Description
<i>sampling_id</i>	integer	The sampling event identifier
<i>date_from</i>	date	The first date (year, month, day) and time (hours, minutes) of the sampling event
<i>date_to</i>	date	The final date (year, month, day) and time (hours, minutes) of the sampling event
<i>varroa_count</i>	integer	The number of varroa mites found in the sampling event
<i>quality_control</i>	integer	<p>The quality level of the sample collected</p> <p>2 = examined with the BeeVS diagnostic system</p> <p>1 = examined manually by a trained group.</p> <p>0 = examined manually by untrained individuals</p>
<i>hive_id</i>	varchar(4)	The hive identifier
<i>yard_id</i>	varchar(4)	The yard identifier

Entity 5: Description of the attributes in the *station* file

Attribute Name	Format	Description
<i>station_id</i>	varchar(6)	The NOAA weather station identifier
<i>station_title</i>	varchar(45)	The NOAA weather Station Name
<i>latitude</i>	decimal(5,3)	Latitude coordinates of the station in decimal degrees in WGS84 standard.
<i>longitude</i>	decimal(6,3)	Longitude coordinates of the station in decimal degrees in WGS84 standard.
<i>station_elevation</i>	decimal(5,1)	Meters above Sea Level

Entity 6: Description of the attributes in the ***weather*** file

Attribute Name	Format	Description
<i>station_id</i>	varchar(6)	The NOAA weather station identifier
<i>date</i>	date	Date (month, day, year) of weather recordings
<i>hour</i>	time	Hour of weather recordings
<i>air_temp</i>	decimal(6,1)	Mean temperature during the hour in .1 degrees Celsius
<i>dew_point</i>	decimal(6,1)	Mean dew point for the hour in .1 degrees Celsius
<i>pressure</i>	decimal(6,1)	Air pressure relative to inches in mercury

<i>wind_dir</i>	decimal(6,1)	The angle between true north and the direction the wind is blowing measured in angular degrees
<i>wind_spd</i>	decimal(6,1)	The rate of horizontal travel of air past a fixed point per hour .1 meters per second

<i>sky_cond</i>	varchar(5)	<p>Code that denotes the fraction of total cloud or other obscuring phenomena coverage.</p> <p>0: None, SKC or CLR</p> <p>1: One okta - 1/10 or less but not zero</p> <p>2: Two oktas - 2/10 - 3/10, or FEW</p> <p>3: Three oktas - 4/10</p> <p>4: Four oktas - 5/10, or SCT</p> <p>5: Five oktas - 6/10</p> <p>6: Six oktas - 7/10 - 8/10</p> <p>7: Seven oktas - 9/10 or more but not 10/10, or BKN</p> <p>8: Eight oktas - 10/10, or OVC</p> <p>9: Sky obscured, or cloud amount cannot be estimated</p> <p>10: Partial obscuration</p> <p>11: Thin scattered</p> <p>12: Scattered</p> <p>13: Dark scattered</p> <p>14: Thin broken</p> <p>15: Broken</p> <p>16: Dark broken</p> <p>17: Thin overcast</p> <p>18: Overcast</p> <p>19: Dark overcast</p>
-----------------	------------	---

<i>precip_1hr</i>	decimal(6,1)	The amount of liquid precipitation measured in .1 centimeters over a one hour accumulation period.
<i>precip_6hr</i>	decimal(6,1)	The amount of liquid precipitation measured in .1 centimeters over a one hour accumulation period

Acknowledgments

The authors wish to thank the Department A10 (Agriculture and Forestry) of the Styrian Provincial government, the Styrian Beekeeping Association, the Austrian Agency for Health and Food Security (AGES), The Rural Training Center (LFI), and The Austrian Beekeeping Federation (Bienen Österreich - Imkereidachverband) for their contribution to the realization of the Austrian Varroa Alert Service and its prototype.

Conflict of Interest

Michel Rubinigg owns the company that produces the software, which is used by beekeepers to provide Varroa infestation data in their bee yards. The data has been anonymized and the coordinates, zip, and lau codes were removed from most of the bee yards in order to hide information that could pinpoint specific beekeepers. This is in line with Terms and Conditions and Privacy Policies for users of the software.

References

- [1] Bees Health (2020) Varroa warning service. <https://bienengesundheit.at/>. (Accessed April 2020).
- [2] Kim Bjerger, Carsten Eie Frigaard, Peter Høgh Mikkelsen, Thomas Holm Nielsen, Michael Misbih, Per Kryger 2019 A computer vision system to monitor the infestation level of *Varroa destructor* in a honeybee colony. Computers and Electronics in Agriculture 164: 1048983. <https://doi.org/10.1016/j.compag.2019.104898>
- [3] NOAA Climatic Data Center Climate Data Online, 2020. <ftp://ftp.ncei.noaa.gov/pub/data/noaa/>. (Accessed December 2020).
- [4] Rosenkranz P, Aumeier P, Ziegelmann B 2010 Biology and control of *Varroa destructor*. Journal of Invertebrate Pathology 103, S96–S119.

- [5] Dietemann, Vincent, et al. "Standard methods for varroa research." *Journal of apicultural research* 52.1 (2013): 1-54.
- [6] Boigenzahn C, Rubinigg M, Wallner T 2020 Der Österreichische Imkereisektor 2019. Biene Österreich - Imkereidachverband: 11.
- [7] Goal 15 | Department of Economic and Social Affairs. (n.d.). Retrieved December 2, 2020, from <https://sdgs.un.org/goals/goal15>
- [8] Goal 2 | Department of Economic and Social Affairs. (n.d.). Retrieved December 2, 2020, from <https://sdgs.un.org/goals/goal2>
- [9] Biodiversity and ecosystems | Department of Economic and Social Affairs. (n.d.). Retrieved January 13, 2020 <https://sdgs.un.org/topics/biodiversity-and-ecosystems>

Predicting Varroa Data & Analytics For Good
Journal External v4.pdf

543
KB