This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2877352, IEEE Transactions on Big Data

Visual Analysis of Spatio-Temporal Event Predictions: Investigating the Spread Dynamics of Invasive Species

Daniel Seebacher, Johannes Häußler, Michael Hundt, Manuel Stein, Hannes Müller, Ulrich Engelke, *Senior Member, IEEE* and Daniel Keim, *Member, IEEE*

Abstract—Invasive species are a major cause of ecological damage and commercial losses. A current problem spreading in North America and Europe is the vinegar fly Drosophila suzukii. Unlike other Drosophila, it infests non-rotting and healthy fruits and is therefore of concern to fruit growers, such as vintners. Consequently, large amounts of data about the occurrence of D. suzukii have been collected in recent years. However, there is a lack of interactive methods to investigate this data. We employ ensemble-based classification to predict areas susceptible to the occurrence of D. suzukii and bring them into a spatio-temporal context using maps and glyph-based visualizations. Following the information-seeking mantra, we provide a visual analysis system *Drosophigator* for spatio-temporal event predictions, enabling the investigation of the spread dynamics of invasive species. We demonstrate the usefulness of our approach in three use cases and an evaluation with more than 30 domain experts.

Index Terms—Geographic information systems, Data visualization, Supervised learning

1 Introduction

Non-native plants, fungus or animal species that out-compete native species often cause severe economic and ecological damage to our planet. With increasing globalization through trade and travel routes, humankind has created opportunities for invasive species to establish themselves in new regions all over the earth.

An exemplary invasive insect currently spreading around Europe and North America is the Asian vinegar fly *Drosophila suzukii* or spotted wing Drosophila (D. suzukii). In 2008, first occurrences were reported in California, Spain, and Italy rapidly followed by other regions and countries [1], [2]. In contrast to other Drosophila species, D. suzukii infests even non-rotting and healthy fruits. It has a wide range of possible host plants that have thin-skinned fruits, like cherries, berries or grapes. An adult female fly can lay 1-10 eggs per fruit and 200-400 eggs within its lifespan of 8-25 days. Depending on temperature and other external factors, these eggs become adult flies within 11-24 days. Thus, 13-15 generation cycles are possible during one year. As a result of the spread of D. suzukii, the USA, for example, noted an annual loss of $500 million [3] in fruit production within a few years. *Agroscop*, the Swiss center of excellence for agricultural research, has also published data on crop losses from 2014 [4] showing that in some Swiss cantons, 80-100% of cherries were unmarketable. Consequently, industry and science are tirelessly searching for novel ways to keep the spread of D. suzukii under control through a better understanding of their spread behavior. Institutes such as the European and Mediterranean Plant Protection Organization[1] or the State Viticulture Institute (WBI) in Freiburg run global databases with weekly to monthly reports about present threats and new findings. Focused on the data gathering aspect these systems are, however, often analytically limited to providing simple D. suzukii distribution maps. To this end, various approaches have been proposed to explore the recorded data. Wiman et al. [5], e.g., make use of the fact that insects are ectotherms, which means that their body temperature equals the ambient temperature. Therefore, low temperatures are a key cause of insect overwinter mortality. The authors tried to estimate D. suzukii populations in different life stages, based on average daily temperatures of some specific fruit production sites combined with trap catches and fruit infestation counts. With their temperature model they found some confirmation of population trends with trap data, and to a limited extent with fruit infestation data. Building on top of this work, other proposed approaches try to optimize temporal and spatial dislocation of control measures by conducting studies on D. suzukii's plasticity of cold tolerance and its overwinter behavior [6], [7]. Spatial and temporal dislocation is caused by mainly measuring in high ripening seasons and at orcharding sites. Focusing on temperature alone neglects the environmental aspects under which the fly could best procreate, or survive even in colder seasons. Other approaches focus on several integrated pest management (IPM) strategies instead. An extensive review of current methods, as well as a categorization, is given by Haye et al. [8]. They introduce strategies that focus on chemical, cultural [9], [10] or biological control [11], [12].

The multitude of approaches shows that analyzing the spread of invasive species is a complex problem. There are many different external influences, which affect the spread of D. suzukii, such

- *D. Seebacher, J. Häußler, M. Hundt, M. Stein and D. Keim are with University of Konstanz.*
  E-mail: firstname.lastname@uni.kn
- *H. Müller is with the State Office for the Environment, Measurements and Nature Conservation of the Federal State of Baden-Württemberg (LUBW)*
  E-mail: Hannes.mueller@lubw.bwl.de
- *U. Engelke is with the Commonwealth Scientific and Industrial Research Organisation - Data61*
  E-mail: ulrich.engelke@data61.csiro.au

Manuscript received XXXX, XXXX; revised XXXX, XXXX.

1. EPPO - https://gd.eppo.int

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2877352, IEEE Transactions on Big Data

as surrounding areas, time, temperature, food supply and many more. This is aggravated by the fact, that these influences have to be considered in a temporal and geospatial context. This illustrates the need of researchers for analyzing large amounts of complex empirical evidence interactively, to gain insights.

In this paper we present our application *Drosophigator* (*Drosophila Investigator*). We follow a visual analytics approach for interactive exploration of large amounts of heterogeneous data sources, including trap counts of D. suzukii, surrounding high-detail land use data, and related metadata. To help researchers investigate the spread dynamics of invasive species, we proceed as follows: Firstly, we interviewed experts to assess their needs, which we then translated into requirements for our presented system. Furthermore, they assisted us in the selection of suitable data sources, which we used to train an ensemble of classifiers to predict time and place of possible occurrences by D. suzukii. These events, the occurrences of D. suzukii, are cumulatively visualized with a glyph-based visualization and brought into a spatio-temporal context by placing them on a map. By allowing zoom and filter capabilities, as well as details on demand, our application enables domain experts to understand the spread dynamics of invasive species. We demonstrate the usefulness of *Drosophigator* in three use cases. Additionally, we collected feedback regarding our application from 37 domain experts. Finally, we discuss our application and highlight potential future work.

2 Related Work

In this section, we first provide an overview of related work for the analysis and prediction of the spread dynamics of invasive species with a focus on D. suzukii. Subsequently, we discuss related work in the visualization of spatial, temporal and spatio-temporal event predictions.

2.1 Predicting the Spread Dynamics of D. Suzukii

Various reviews of methods for the prediction of the geographic expansion of D. suzukii have been introduced. Cini et al. [1] argue that while modeling spread dynamics seems to be an important first step in understanding the population dynamics of D. suzukii, the consideration of host plant effects, such as host plant species phenology and density, has to be a research priority for future work. In another work, Asplen et al. [3] provide in-depth information about D. suzukii and propose a general research agenda for future pest management.

As a crucial starting point, they consider the monitoring of D. suzukii to collect and identify the data which are necessary for the prediction of the spread dynamics of invasive species. Consequently, several projects focus on the monitoring of invasive species such as VitiMeteo [13] or Drosomon [14]. As pointed out by Asplen et al., further research is now needed to develop various pest management tools and to facilitate the transfer of the generated knowledge to users. Information visualization has shown to be effective in this regard, since it is the communication of abstract data through the use of interactive visual interfaces [15].

2.2 Visualization of Spatial, Temporal and Spatio-Temporal Data

When analyzing the spread dynamics of invasive species, adequate visualization techniques are required to incorporate the spatio-temporal aspects of the available data. To this end, Andrienko et al. provide an overview [16] about existing exploratory techniques related to spatio-temporal data and the corresponding tasks. Spatial event distributions as well as predictions are often visualized with the help of a map [17], [18], heat maps [19], [20], [21] or choropleth maps [22], [23]. Glyph-based visualization for geographical topic comparison have been introduced as another way of analyzing contextual spatial data [24]. Their use has been demonstrated by analyzing Twitter and news stream data to detect and visualize important discussion topics on a map, illustrating topic distributions for different countries.

The visualization of temporal data is still a common discipline in the information visualization community. The state-of-the-art evolves around traditional visualization techniques such as line graphs. For periodical temporal data (e.g., hourly, daily or monthly readings), circular visualizations have increased in popularity over the last decade, as Fuchs et al. [25] showed, that the radial encodings of time are more effective when a user has to pick particular temporal locations. This includes, for example, dense-pixel displays [26] or spiral visualization techniques [27], [28].

Multiple coordinate views have often been proposed [29] to visualize both geographic as well as temporal data. However, combining both aspects into a single visualization is more desirable, since this reduces the cognitive effort for the user [30]. One approach that combines the advantages of circular visualizations for periodic temporal data, forming a single visualization, are RingMaps [31].

2.3 Positioning of our work

Research on invasive species which conquer new environments is characterized by the fact that distribution processes are unknown and data is sparse. We account for this key characteristic in our analysis method and include the visualization of derived uncertainty and statistical importance measures. We propose a single visualization of the spatial and temporal dimensions of predictions of the spread of D. suzukii, using maps and circular glyph-based visualization. Additionally, we extend this approach by allowing multiple event types, including the uncertainty of the prediction in the visualization. Our work is novel in that we combine our glyph visualization with state-of-the-art information visualization and interaction techniques to enable experts to seamlessly analyze micro- and macroecological factors regarding the spread dynamics of invasive species, with the example of D. suzukii.

3 Data Description

We performed several expert interviews with the State Viticulture Institute (WBI) in Freiburg, Germany, in order to gain a better understanding of the influences and factors about the spread of D. suzukii as well as to identify current challenges faced by domain experts. The WBI offers, through their web service *VitiMeteo*, forecast models for different fungi species, monitoring data for various pests, as well as weather data related to viticulture in the federal state of Baden-Württemberg. In our interviews we found that although a lot of data about D. suzukii is being collected by the WBI, they lack adequate methods to analyze and interpret the increasing amounts of data as well as visualization techniques to communicate and present related findings.

In the data provided by VitiMeteo are, among other things, observations of the spread of D. suzukii. This data consists of trap findings of D. suzukii as well as percentage information about

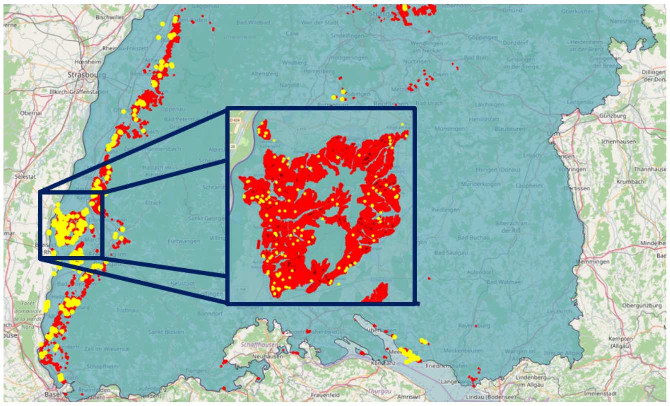


Fig. 1: Vineyards (highlighted in red) in Baden-Württemberg, as well as measurements stations by the WBI (highlighted in yellow). Highlighted is the Kaiserstuhl, one of the biggest wine regions in Baden-Württemberg.

how many berries were infested in a sample taken at the station. Additionally, there is percentage information about how many eggs were found in a sample. This percentage can be over 100 % if there are more egg findings than berries in a sample. These observations are collected from 867 stations non-uniformly spread over Baden-Württemberg as shown in Figure 1. Some of them only report observations for one day, others report multiple observations over a time period of up to 1641 days. The observations are rather sparse and irregularly sampled, which makes the use of standard time series analysis techniques challenging, if not impossible. Consequently, Drosophigator should enable researchers of the WBI to interactively analyze this complex data source.

The Julius-Kühn-Institute [32] (JKI) suggests that the number of trap findings are increasing in late summer and stay high until winter. Pelton et al. [33] found that areas surrounded by woodland exhibit an earlier infestation. Additionally, as highlighted in the related work, Asplen et al. [3] recommend considering host plan effects, such as host plant density. As a result, the focus of our application is the analysis of the spread dynamics, exemplified by D. suzukii, by taking temporal distribution as well as environmental factors into account. In order to test the hypotheses of the JKI and Pelton et al., we gathered the relevant data from different resources. The time of year is already present in our observation data provided by the WBI. To gather the height of every measuring station, we make use of the ASTER Global Digital Elevation Map [34] which was released by the Ministry of Economy, Trade, and Industry (METI) of Japan and the United States National Aeronautics and Space Administration (NASA). Information on land use and land cover for the state of Baden-Württemberg was derived from the ATKIS [35] and ALKIS [36] datasets created by the State Agency for Spatial Information and Rural Development Baden-Württemberg. This includes high-detail statewide land use information. It consists of main groups, such as forests or industry, but also subgroups, such as coniferous forest or treatment plans. Overall there are 84 different combinations of groups and subgroups.

In interviews with experts, we found out that the climate is also a relevant factor, which should be considered in the analysis of the spread dynamics of D. Suzukii. Based on these findings, we have extended the data set to include meteorological data. The meteorological data was provided by the German Meteorological Office and provides information on various features such as hours of sunshine, cloud cover, wind strength and direction, as well as temperature and precipitation at over 300 stations. However, many of these features have severe data gaps, so for subsequent analysis, only the precipitation and temperature features are used, resulting in 7 additional features for our feature vector.

4 Ensemble-based classification of infested areas

To identify regions, in our case vineyards in Baden-Württemberg, which are potentially endangered by D. suzukii, we use machine learning to train a model using the data provided by the WBI in combination with the data collected from ATKIS and ASTER. This allows us to learn which combination of features make areas, at certain points in time, susceptible to the occurrence of D. suzukii. By applying the trained model on other areas we can find new potentially endangered areas.

4.1 Data Preparation

To training our model we need to determine which areas are severely affected by D. suzukii and which are not. As mentioned in Section 3, we have three types of observations (trap findings, berry infestation and egg findings) which all indicate whether D. suzukii occurs in a specific area. All of these observations serve as indicators that an area is susceptible to the occurrence of D. suzukii, thus allowing us to combine them into a single measurement by first normalizing them to the range $[0, 1]$ and afterwards summing them up into a single feature, subsequently referred to as *observations*. To cope with irregular samplings of measurements, we averaged the number of observations per station per month. The resulting distribution is right-skewed, with most values being 0, meaning that for most stations we observe no occurrence of D. suzukii in a month. To still be able to differentiate between stations with a high occurrence of D. suzukii and stations with only low or no occurrence, we decided to set the 70 % percentile as an experimental threshold to classify our stations. This threshold may be changed later, requiring a retraining of our model, but otherwise not affecting the later steps of the classification and the usage of our application. In total we have a training set consisting of 7224 instances. Using the 70 % percentile of the average *observations* per month to partition our data into low occurrence (negative) and high occurrence (positive) areas gives us a data set with 1650 positive and 5574 negative instances.

We enriched these instances, by adding information about the environmental surroundings of each station. First, we added the height information, which we extracted from ASTER. Second, we added the surrounding land use information. Since a local spread is possible by D. suzukii itself, we extracted the land use information in a 5 km radius around each station. In addition, we have included seven weather features, for which we have extracted the data of the nearest weather station. Finally, we have an 91 dimensional feature vector for each instance, consisting of the station height, the surrounding land use, and the meteorological data.

Using this partitioning we end up with a rather imbalanced data set with over three times as many negative examples as positive ones. This can cause problems since many machine learning algorithms depend on the assumption that the given training data set is balanced [37]. Although machine learning techniques exist which can deal with imbalanced data sets, such as the *Robust*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2877352, IEEE Transactions on Big Data

Decision Trees of Liu et al. [37], we want to employ ensemble-based classification, which is a combination of different classifiers. This allows us to improve the classification performance [38] and to model the uncertainty of our classification, which aids people in making more informed decisions [39]. This requires the creation of a balanced data set, which we can achieve by either using undersampling of the majority class or oversampling of the minority class. Undersampling can be achieved through stratified sampling using the occurrence class as strata. However, this would remove instances from our already small data set. To avoid this, we employ oversampling of the minority class using the Synthetic Minority Over-sampling Technique (SMOTE) [40]. SMOTE picks pairs of nearest neighbors in the minority class and creates artificial instances by randomly placing a point on the line between the nearest neighbors until the data is balanced. Thus, allowing us to employ default machine learning algorithms.

4.2 Ensemble-based Classifier Training

For training the classifiers we use the state-of-the-art data mining systems KNIME [41] and WEKA [42]. We use a selection of well-known machine learning techniques such as Decision Tree, Random Forest, Multilayer Perceptron, k-nearest neighbor classifiers, etc. This selection was determined in an experimental evaluation of available classifiers in KNIME and WEKA, and might be extended later. In order to support our decision to employ ensemble-based classification to improve the classification performance, we first need a baseline measurement. We performed a 10-fold cross validation of each of the classifiers mentioned in the previous paragraph and found that the Random Forest classifier achieved the best performance, with a mean Cohen's κ score of 0.659. The other classifiers achieved Cohen's κ scores between 0.188 and 0.659, as shown in Table 1, which are according to Altman [43] poor to good agreement between the prediction and actual class. To test if ensemble-based classification could achieve better results, we used stacking [44]. Here a logistic regression model is trained which uses the prediction of the previously trained classifiers as inputs to make the final prediction, as suggested by Ho et al. [45]. Using this approach we achieved a Cohen's κ score of 0.701, which is better than all the individual classifiers that were tested in this evaluation. Additionally, we are now able to model the uncertainty of our prediction, which according to Skeels [39] is important for decision-making.

Classifier	**Cohen's κ**
Ensemble-based Classification	0.701
Random Forest	0.659
Decision Tree	0.569
1-NN Classifier	0.559
K*	0.534
2-NN Classifier	0.439
4-NN Classifier	0.436
Multilayer Perceptron	0.422
3-NN Classifier	0.419
5-NN Classifier	0.396
LibSVM	0.368
Kernel Logistic Regression	0.362
Bayesian network	0.34
Naive Bayes	0.188

TABLE 1: The ensemble-based classification achieved the best results, in accordance with the study by Rokach [38]

4.3 Feature Importance

As shown in the last section, ensemble-based classification, either with stacking or with the random forest classifier, achieved the best results in our evaluation. Additionally, ensemble-based classification has the added benefit of allowing us to model the uncertainty of our prediction. These benefits, however, come with the drawback of lacking interpretability. For individual predictions, it is no longer possible to determine which combination of features is responsible for the obtained result. However, it is possible to determine the global influence of individual features. While this still does not allow us to explain individual predictions, it allows us to highlight the most important features. Thus, even with high-dimensional data, we can quickly guide users to the most important features so that they can make an informed decision when comparing predictions for multiple areas.

To calculate the importance of each feature, we measure its impact on the evaluation result. We iteratively choose one feature $f \in F$ of the existing features and shuffle its values randomly, while leaving the remaining features $F \setminus \{f\}$ untouched. This removes any relationship between the selected feature and the output class. Then we perform a 10-fold cross-validation, as described in the last section, to measure how the shuffling of a feature affects the performance of the classifier. If the influence of the feature is large, the classification result should deteriorate considerably, but if the influence of a feature is small, the classification result should remain the same. Therefore, we define the importance of a feature f as $1-$Cohen's κ, where Cohen's κ is the performance of the classifier on the data set with shuffled variable f. Table 2 shows the calculated importance, normalized to a scale of $[0-1]$ for a selection of features. A similar approach to measure variable importance was presented by Breiman [46], which calculated the increase in the misclassification rate of random forests when permutating a random variable. *Precipitation* is the most important feature, followed by *altitude*, *minimum temperature*, as well as the land use features *closed building areas*, *hedges and shurbs*, *meadow orchards and farmland*. The least important features include *fairgrounds*, *traffic*, as well as *rest areas*.

Feature	**Normalized Importance**
Monthly Precipitation	1
Altitude	0.805
Max. Daily Precipitation	0.766
Closed Building Areas	0.747
Hedges and Shrubs	0.695
Meadow Orchard	0.676
Min. Daily Temperature	0.676
Farmland	0.642
...	...
Traffic	0.101
Fairground	0.0

TABLE 2: The calculated importance ($1-$Cohen's κ) for individual features. *Precipitation* was is considered to be the most important feature. This is followed by *minimum temperature*, *altitude* and the *surrounding agriculture*, *hedges and shrubs*, and *closed building areas*.

5 DROSOPHIGATOR: VISUAL ANALYSIS OF SPATIO-TEMPORAL EVENT PREDICTIONS

Just providing users with the raw results of our prediction is not sufficient as we generate over 20.000 predictions for all months and vineyards in Baden-Württemberg for the year 2016. Furthermore,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2877352, IEEE Transactions on Big Data

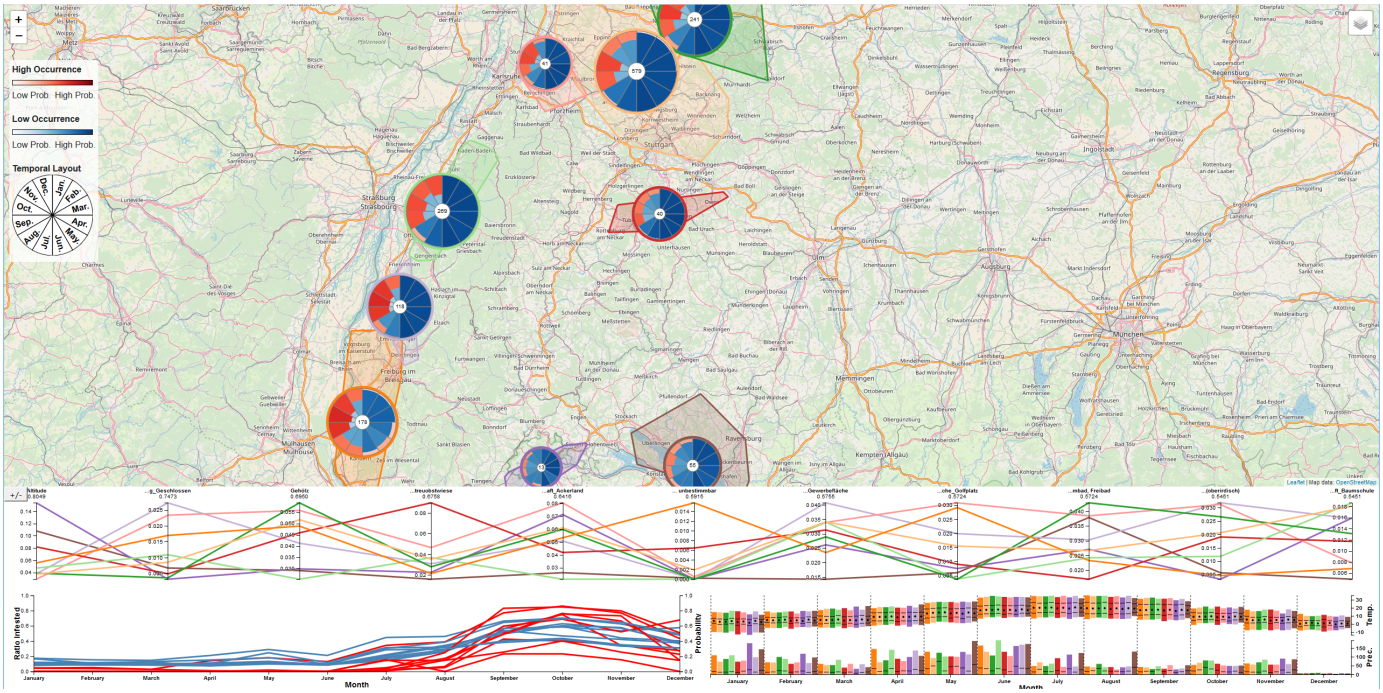


Fig. 2: An overview of the *Drosophigator* application for the visual analysis of spatio-temporal event predictions. Our application is divided up into two parts. An interactive map, including our glyph visualization of the spatio-temporal event predictions and three data visualizations, allowing the investigation of the aggregated environmental and meteorological features and temporal occurrences of D. suzukii.

the raw results do not provide spatial context. Thus, it is not interpretable which makes it hard for experts to integrate their domain knowledge into the analysis process. Hence, we need visualization to help experts to identify spatial and temporal patterns more easily. To achieve this, we follow the visual information seeking mantra of Ben Shneiderman: “Overview first, zoom and filter, details on demand” [47].

To improve our understanding of the needs of domain experts, we performed additional informal expert interviews with members of the Julius-Kühn Institute for Plant Protection in Fruit Crops and Viticulture [2]. We presented a first prototype of our application to five experts in the field of biology, with a focus on invasive species, one of whom with more than 30 years of work experience and asked them for feedback and further requirements to analyze the spread dynamics of D. suzukii. We use the requirements gathered in this interview in the design process of our application *Drosophigator* to tailor its capabilities to the needs of the users. The following list is a summarization of the collected requirements:

V Visualization Requirements

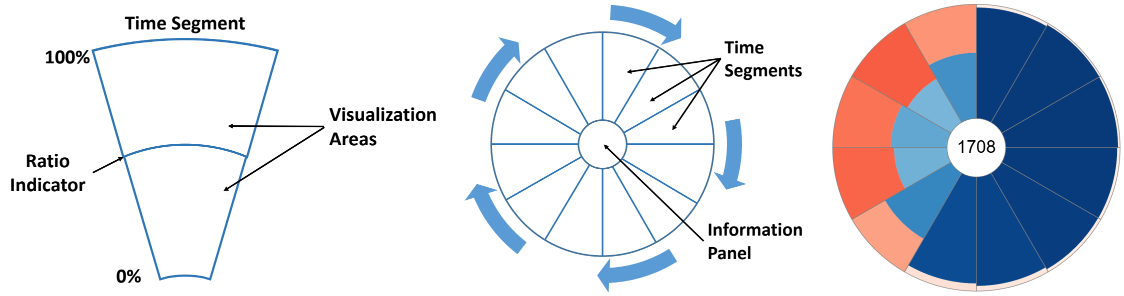
- V1 *What does the geographical context look like?* It is essential for experts to get information about the geographic context of a prediction. This context information enables them to incorporate their expertise. For example, they know the dominant type of wine in certain regions and how susceptible it is to infestation by D. suzukii. Thus, allowing them to better interpret and validate the results of the prediction.
- V2 *When does the infestation happen?* The temporal occurrence of D. suzukii is very interesting for domain experts. Visualizing this information helps to identify interesting or unusual patterns and is necessary when investigating certain hypotheses, such as the effect of the surrounding environment on the temporal occurrence of D. suzukii.
- V3 *How to inspect individuals and aggregates?* For experts both the visualization of the prediction for individual vineyards and aggregated regions are interesting. This enables them to investigate *micro-ecological* differences such as environmental characteristics of an individual vineyard, but also *macro-ecological* effects, such as the dominant type of wine in a certain region, since different types are affected differently by D. suzukii.
- V4 *How certain is the prediction?* The visualization of prediction uncertainty is of high relevance for biologists. Knowing how certain the prediction algorithm is for specific regions allows them to either investigate or filter out regions with a high uncertainty.

I Interaction Requirements

- I1 *How can I see details-on-demand?* The experts stressed the importance of getting details-on-demand. When analyzing the occurrence of D. suzukii, for instance for a single vineyard, they need to get further information such as the exact location and outline of the vineyard, the surrounding environment, and the infestation.

2. https://www.julius-kuehn.de/en/plant-protection-in-fruit-crops-and-viticulture/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2877352, IEEE Transactions on Big Data



(a) Each time-segment can be divided into multiple visualization areas, which correspond to the number of possible events. The more interesting events to observe are placed on the outside, to be more easily visible. The ratio indicator is used to illustrate which event is most likely to occur.

(b) Time-segments are ordered clockwise. They can represent different time units, from hours, days, months or years. Each time-segment can be used to visualize the event prediction for the specified time frame.

(c) Example Glyph Visualization showing the ratio of vineyards with a high occurrence of D. suzukii and vineyards with a low occurrence, as well the uncertainty of the prediction for one year.

Fig. 3: Visual explanation and example of the designed glyph showing the prediction results for each month of a certain area, (a) Sketch of a single time-segment, (b) the resulting glyph-based temporal event prediction visualization and (c) a real glyph example.

I2 *How can I compare regions?* The experts want to be able to compare multiple regions to find out differences between the independent variables, such as environmental features and the dependent variable, i.e., the prediction of D. suzukii.

To satisfy requirement **V1**, we build a geographic information system, using a map as the basis for interaction and spatio-temporal analysis. We consider the familiarity of domain experts with this kind of visualization as an additional benefit. We visualize our predictions on the corresponding position on the map so that users are immediately aware of the geographic context. Additionally, combining our geographic visualization with a visualization of the temporal predictions into a single visualization is more effective, since this requires less cognitive effort for the users, than mentally linking multiple views [30].

With respect to requirement **V1** and to satisfy requirement **V2**, we enrich our map visualization with a temporal visualization of the occurrence of D. suzukii. Existing related systems such as BirdVis [48] offer heat map overview visualizations. However, as we want to investigate the distribution of a species over time, we designed a map overlay consisting of several glyphs. This partially preserves the geographic context while the glyphs can be used to encode additional contextual information. Extensive work on glyphs has been done in the past which we used as guidelines to design the final glyph proposed in this paper. This includes, e.g., the work by Fuchs et al. [25] and Borgo et al. [49]. The goal of our glyph is to visualize whether a certain region is endangered or not. Consequently, we visually encode the classification results of a specific month represented by its time segment. The basic design of a time segment is depicted in Figure 3 (a). Therefore, we make use of the interior of the respective time segment to represent the classification results of the ensemble-classifiers applied in Section 4. For each month we have a distribution of safe and endangered vineyards, according to the classification. Since the number of vineyards stays the same over all months for each glyph, we fill the area of the time segments according to the ratio of the binary outcome (endangered, not endangered). This technique results in a radial glyph similar to a stacked bar chart showing fractions of the whole. We use the colors red (endangered) and blue (not endangered), as derived from the warm-cold color scale [50] to distinguish the outcome.

As we, additionally, are aware of the probability of each outcome (endangered or not), we can use this to represent the (un-)certainty from **V4** and to, e.g., find out where to place additional measurement stations. The average probability of a given outcome is encoded using the respective half of the warm-cold color scale, such that a high probability/certainty results in a stronger color tone while a low probability, on the other hand, is represented by a weaker color tone. An intense red color, for instance, means that there is a very high probability of endangered vineyards within the respective month (time segment) and area (glyph location). We have included a legend, as shown in Figure 2 or Figure 6, which provides the user with more information about the used colors and the temporal layout of the time segments. Additionally this technique enables the user to spot potentially new measurement areas, by detecting areas with a high uncertainty value. An overview of the realized glyph representing all vineyards in Baden-Württemberg can be seen in Figure 3c. It can be observed that there is a general trend as the number of endangered vineyards (red) is increasing rapidly in late summer and stays high until early winter. This observation corroborates the hypothesis that D. suzukii may only survive in a relatively stable environment regarding temperature such that it dries out in the summer and freezes in winter months.

The resulting glyphs are positioned on their on top of the respective vineyards in the map visualization. However, to fulfill requirement **V3**, we need to provide an aggregate visualization for whole wine regions. Hence, we use clustering to group vineyards, depending on the zoom level. The farther we zoom out, the more vineyards are clustered together. In order to give the users an indicator for the number of grouped vineyard in each glyph, we scale the radius of the glyph on the one hand, and on the other hand give exact information about the number of vineyard areas in the information panel of the glyph. For the dynamic aggregation based on the zoom-level, we use the *markercluster* library [51], which has a greedy hierarchical clustering algorithm and allows real-time joining and splitting of clusters, even with up to 50.000 points. This approach is scalable to very large problems. PruneCluster [3] is an alternative library which uses a more performant clustering algorithm, which can scale this approach up to a million points. This allows for a seamless investigation

3. https://github.com/SINTEF-9012/PruneCluster

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2877352, IEEE Transactions on Big Data

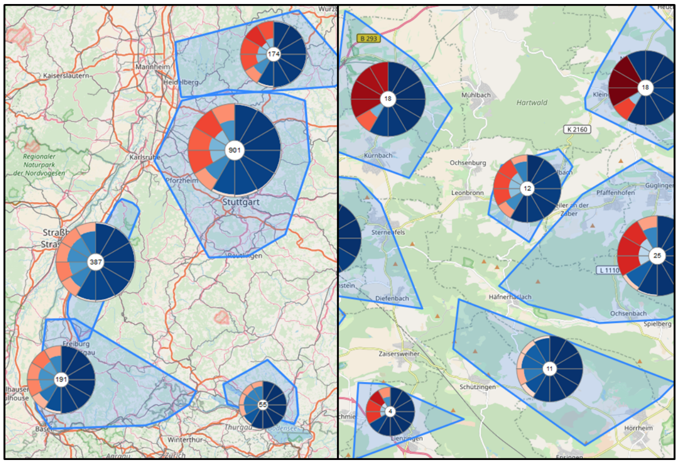


Fig. 4: Overview of our glyph-based visualization. For each cell, the predictions and their uncertainties are averaged per time-segment and visualized in our glyph. We provide zoom-and-filter capabilities by allowing the user to zoom in and out of the map.

of differently sized regions, depending on the desired analysis use-case, as shown in the example in Figure 4. On the left side, all vineyards in Baden-Württemberg are clustered into five groups, with up to 900 vineyards in a single group. The convex-hull of all vineyards contained by a cluster is highlighted in blue in the background. On the right side of Figure 4, a more fine-grained clustering with five to twenty vineyards per group can be seen.

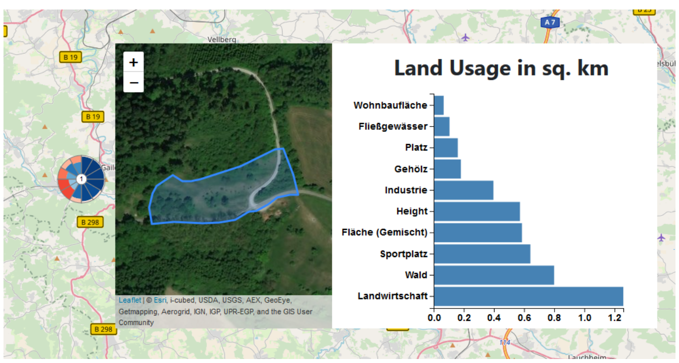


Fig. 5: Overview of our details-on-demand visualizations: We provide the user with details on the structure and environment of the vineyards by highlighting the outline of the vineyard in a satellite image. We show the absolute area of the surrounding land use in a bar chart visualization. Additionally, we provide a view with the relative percentage distribution of the environmental features compared to other vineyards in the viewport and visualization of infestation information and uncertainty of our prediction.

During our expert interview, the experts stressed their need for details-on-demand visualizations, when investigating vineyards, such as the outline of the vineyard and the surrounding land use.

To accordingly satisfy requirement **I1**, we added semantically meaningful tooltips for individual glyphs, as well as three data visualizations, a parallel coordinates plot (PCP) of the environmental features, a box plot inspired visualization of the climate data, and a line graph (LG) of the infestation and uncertainty. An overview of these visualizations is shown in Figure 2. By default, only 10 dimensions are visualized in the PCP, since a visualization of all 84 land use features would lead to overplotting. We make a preliminary selection of the most important land use features, with the help of the feature importance, which we calculated in section 4.3. However, the user can add or remove features manually. To assist the user in the selection, he can inspect the importance of each feature before it is added to or removed from the PCP. In our tooltip, we highlight the vineyard in a satellite image, which enables the domain experts to more closely investigate the outline and the surrounding land use. Additionally, we provide a bar chart of the most prominent environmental features in the surrounding area in absolute values.

In the PCP we also provide information about the land use using relative measurements. The axis of the PCP provide information about the percentage distribution of the environmental features of a region or vineyard, meaning that summing the values of one line over all axis results in 100 % land use. Each axis is scaled to the minimum and maximum value of a particular environmental feature of all vineyards in the current viewport. This prevents that the scale of features with a low percentage of the overall land use are dominated by larger areas. Furthermore, this enables experts to compare the environmental features of the vineyards currently in the viewport as requested in requirement **I2**. An example for the comparison of four wine regions regarding their environmental features is shown in Figure 7. Additionally, we provide *brushing and linking* capabilities between all visualizations. By hovering over a glyph visualization or by selecting multiple glyphs, their respective environmental features, as well as infestation and uncertainty measures, are highlighted in the PCP and LG. Vice-versa hovering over a line in the PCP or LG highlights this instance in our other visualizations. PCP and LG also support brushing, allowing the user to filter out instances. This enables experts to focus on desired vineyards, for instance, vineyards with a high infestation in April and May or vineyards with few industrial sites in their vicinity.

6 Use Cases

In this section, we highlight how visualization can help domain experts to gain insights about the spread dynamics of D. suzukii. For this purpose, we demonstrate how experts can use the interactive map in combination with our glyph-based visualization and the linked data visualizations to investigate complex micro- and macro-ecological factors and influences on the spread of D. suzukii. Therefore, we investigated three recently proposed assumptions about the time of infestation [32] and the influences of environmental [33] and meteorological factors [52]. These use cases will serve as examples of how domain experts without special knowledge of data analysis can use Visual Analytics applications to examine large and complex amounts of data.

The JKI states as a general rule, that the number of findings increases with decreasing temperatures in late summer and stays high until November or later if there are no cold snaps [32]. To investigate this hypothesis we create an overview of all available predictions. For this we employ our semantic zoom to dynamically aggregate all vineyards in Baden-Württemberg into a single glyph. The resulting glyph-visualization is shown in Figure 6. In the visualization, we can easily recognize that the number of infestations is marginal in the first half of the year. However, there is a strong increase in the predicted number of infestations and diminishing uncertainty starting in August until December. Using our weather chart, we can identify, that starting in December, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2877352, IEEE Transactions on Big Data

average daily temperature is dropping below 3°C. This observation is consistent with the hypothesis of the JKI, which states that the mortality rate of D. suzukii increases strongly below a temperature of 3°C.

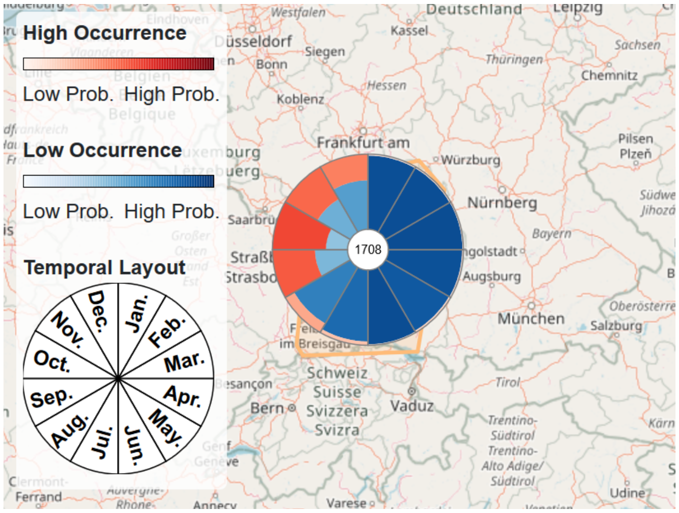


Fig. 6: Overview glyph-visualization of all vineyards in Baden-Württemberg. The development over the time-segments shows that the severity of infestation and the certainty of our prediction increases in late summer and stays high until the end of the year. This corroborates the hypothesis of the JKI [32].

A recent two-year field study of Pelton et al. [33] suggests, that high amounts of surrounding woodland are correlated with an earlier infestation of D. suzukii. By using a finer resolution, we can identify four neighboring vineyards near the city of Heilbronn, which show strong differences in the earliness of the infestation by D. suzukii, as shown in Figure 7. The two bottom-left vineyards (highlighted in green colors) shows an earlier infestation than the two top-right vineyards (highlighted in red colors). To identify differences in the land use of these vineyards, we employ our parallel coordinates plot. We can see that the bottom-left vineyards, which exhibit an earlier infestation, have a larger amount of hardwood in the surrounding area and have a lower altitude than the two top-right vineyards. This may be an indication that in the bottom two areas D. suzukii has more potential natural refuges, while in the top two areas it is more exposed. This finding coincides with the hypothesis of Pelton et al. However, the parallel coordinates plot also gives an indication about the feature importance. Hardwood only has a normalized importance of 0.41, which might be an indicator that this is not the main feature responsible for the difference in the earliness of infestation and that further investigations should be carried out.

In their study on the impact of meteorological factors, Santos et al. [52] identified that one of the most influential factors was the annual precipitation. In order to examine whether this finding also applies to the vineyards in Baden-Württemberg, we take a closer look at the major wine-growing regions, highlighted in Figure 8. We can identify that all wine-growing regions show a medium to large infestation by D. suzukii during late summer, except for the Bodensee region in the bottom right. Using our box plot inspired visualization to investigate the meteorological features, we see that are no major deviations in the temperature. However, in 2016, especially in July and August, the Bodensee region experienced much higher precipitation compared to the other regions. This is in accordance with the results of Santos et al., which say that too little or too much precipitation reduces the infestation of D. suzukii.

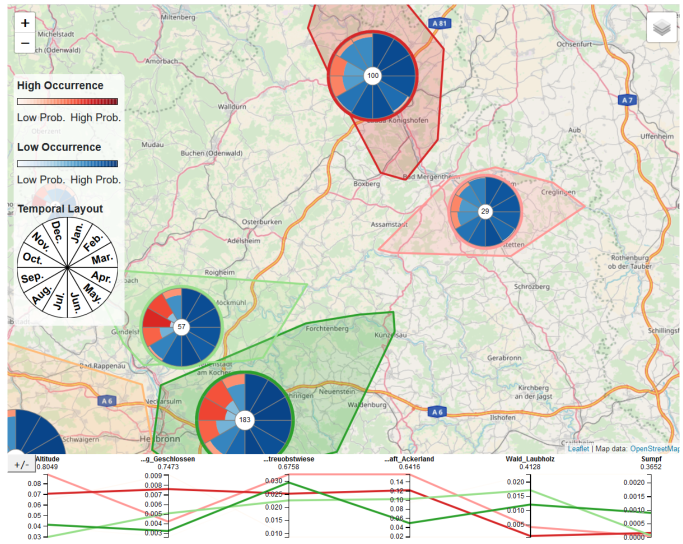


Fig. 7: Comparison of four neighboring vineyards. The bottom-left vineyards (highlighted in green colors) exhibit an earlier infestation by D. suzukii that the top-right vineyards (highlighted in red colors). The parallel coordinates plot shows that the vineyards in the lower cells have more surrounding deciduous forest (*Laubholz*), as well as, a lower altitude, than those in the upper cells. This finding strongly supports the hypothesis of Pelton et al. [33], that forests can act as a natural habitat for D. suzukii, thus enabling an earlier infestation.

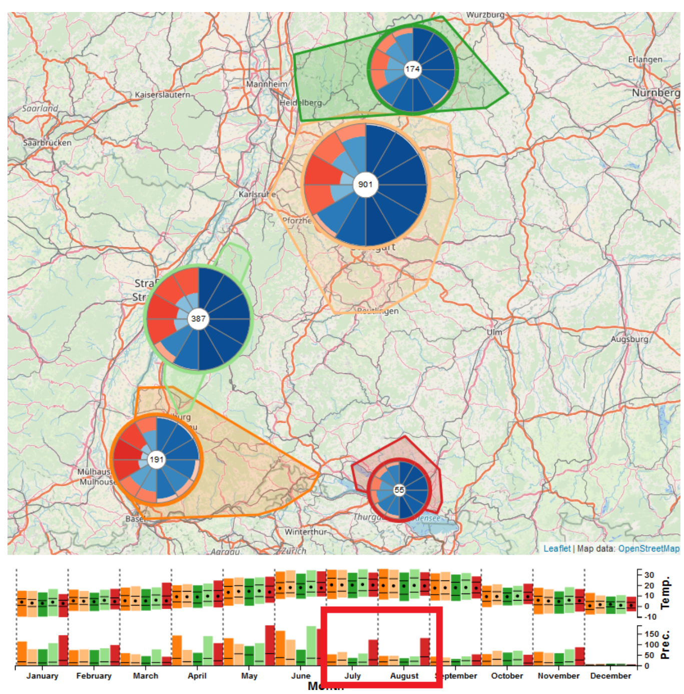


Fig. 8: Comparison of the major wine-growing regions in Baden-Württemberg. The Bodensee region (red) in the bottom right shows a lower infestation by D. suzukii, compared to the other regions. This could be due to the significantly higher precipitation in July and August, which is in line with the results of Santos et al. [52].

In these three use-cases, we have demonstrated the capabilities of our tool. Following the information-seeking mantra of Shneidermann using our glyph-based visualization of the ensemble-based predictions, as well as the uncertainty of the prediction, allows us to make observations supporting hypotheses of researchers about the spread dynamics of D. suzukii.

7 Expert Feedback

To gather feedback from domain experts about our system, regarding the design decisions made and potential improvement ideas, we presented our system at the 6th workshop of the working group "D. Suzukii" [53] on the 5th and 6th of December in Bad Kreuznach, Germany.

The goal of this workshop was the mutual exchange of knowledge between researchers and practitioners. Over 80 biologists, researchers, agri- and horticulturists from various countries participated in the workshop. The focus of our talk was our application *Drosophigator*. We first gave the experts a brief introduction to the area of Visual Analytics and pointed out how this approach can support them in analyzing the spread dynamics of D. suzukii. Afterwards we presented our system in detail and explained the individual components to them. For example, how to interpret the glyph or how to work with PCP in combination with brushing and linking to perform different analysis tasks. In addition, we presented the possibilities of the system on the basis of various use cases. After the presentation of the application, a questionnaire was handed out to the workshop participants, in which they could evaluate the different aspects of our application and give us additional information about their background. The participants had time for the remaining one and a half days of the workshop to fill in the questionnaires, examine our system in detail and ask us additional questions. We used the results of this questionnaire to evaluate our system and design decisions. A limitation of our evaluation approach is that it does not capture the actual practice of working with the tool, but instead provides a first impression of whether this type of system can support the expert in their use cases, since the experts only had a limited time for interaction with the system and also no comparison with other systems. In addition, in order to better capture the first impression of the experts, they were able to provide us with open feedback.

7.1 Questionnaire Design

The questionnaire was designed to capture feedback about the visualization and interaction design and the analysis capabilities of our system, as well as personal information about the participants of the workshop. For the personal information, we asked users to specify their gender, occupation, as well as working experience in years. Additionally, we asked about their computer expertise (Expert to Novice) and frequency of computer usage (daily to never), which could be answered on a 5-point Likert scale [54].

For the system feedback, we asked the users about the various components of our system. Participants could answer via a 5-point Likert scale (Strongly Agree to Strongly Disagree). The questions were formulated in German. The following list provides close translations:

V1 Temporal data is arranged comprehensibly in the glyph.
V2 The glyphs helps me to understand the occurrence of D. suzukii.
V3 The visualization on a map helps me to interpret the results.
I1 The interactions in the presented system are comprehensible.
I2 The offered interaction possibilities are sufficient.
A1 I find it important to have overview visualizations (map of Baden-Württemberg) as well as detail visualizations (single vineyards).
A2 I can infer causes for the occurrence of D. Suzukii from the various visualizations.
A3 The presented approach can be applied on other invasive species.
A4 The presented approach can help me with my work.

7.2 Participants

The questionnaire was filled out by 37 participants (16 male, 20 female, one abstention). 8 participants work in the field of agriculture and 4 in the field of horticulture, there were 12 biologists and 10 researchers or research associates and 3 with other or undisclosed jobs. The mean working experience in years of the participants was 12.64 years, with a standard deviation of 11.70 and a range of $[1, 40]$. Accordingly, we consider the feedback of the participants as highly valuable for the evaluation of our approach.

7.3 Findings

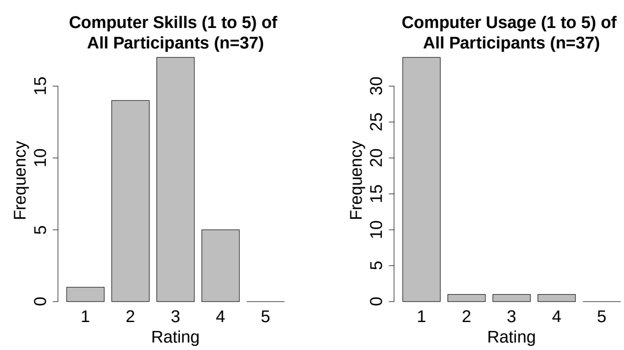


Fig. 9: Evaluation of the computer skills and computer usage habits of all participants (n=37). The participants rated these questions on a scale from 1 (expert, daily) to 5 (novice, never).

The results of our evaluation are very promising. As shown in Figure 9, most of the participants (60 %) stated to be intermediate or novice computer users. This is in stark contrast to the computer usage habits of the participants, where 90 % stated to use the computer daily. This distribution makes the need for intuitive and interactive systems, which support the user, clear.

The evaluation of the system feedback of all participants is shown in Figure 10. Very noticeable are the results for question ***A1***. Over 80 % of the participants state that they find it important to have both overview, as well as detail visualization, confirming our design decision to follow the visual information-seeking mantra of Shneiderman. The feedback with respect to our visualization design (***V1, V2, V3***) is also positive. Nearly two-thirds of the participants stated that temporal data is arranged comprehensibly in the glyph and that the glyph helps them to understand the occurrence of D. suzukii. Additionally, 62 % agree with our decision to combine the abstract visualization of the temporal data with a map, since it helps them to interpret the results of our classification by providing necessary context information such as location.

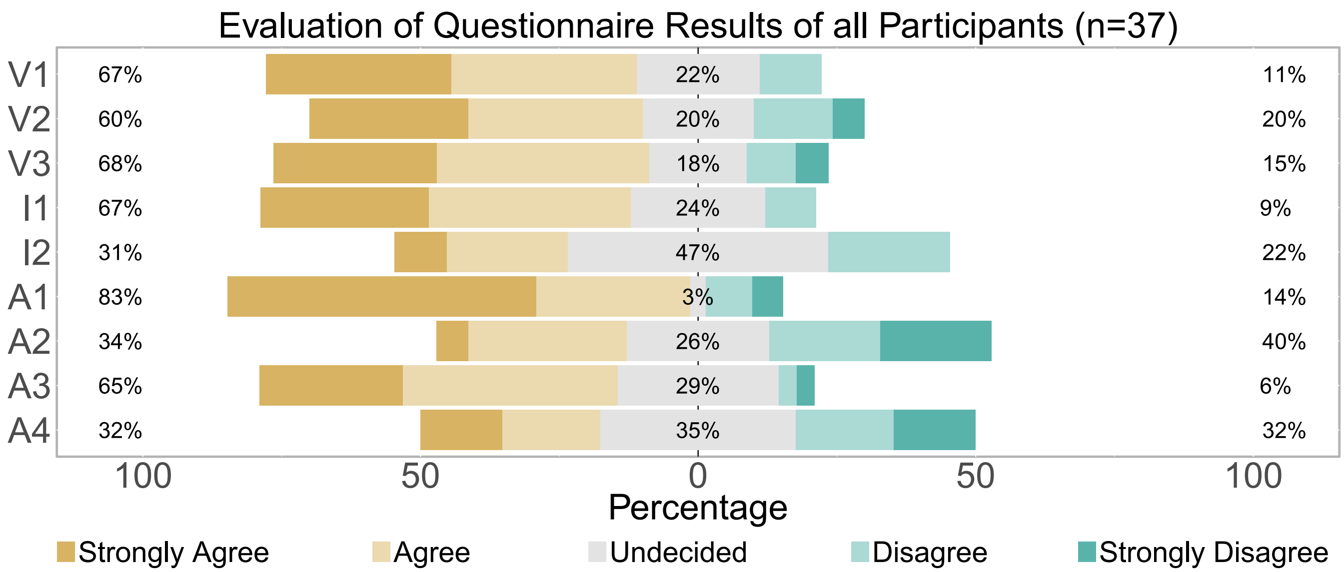


Fig. 10: Evaluation of system feedback of all participants (n=37). Shown are the responses of the participants on questions regarding the visualization design (*V1, V2, V3*), the interaction design (*I1, I2*) and the analysis capabilities (*A1, A2, A3, A4*) of our system Drosophigator in a diverging stacked bar chart visualization [55].

The feedback regarding the interaction design (***I1, I2***) is confirming our design, too. While most of the participants agree that the interaction design is comprehensible, many are undecided whether the interaction possibilities are already sufficient. In our opinion, experts would highly value even more ways to interact with the system, allowing them to change the time-granularity or adjust the classification by providing additional information. This trend is also reflected in the feedback with respect to the analysis capabilities of our system (***A1, A2, A3, A4***). Most of the participants agree that our visualizations are comprehensible, that the decision to follow the visual information-seeking mantra is justified and that this approach can be applied on other invasive species.

Besides the questionnaire, the experts also had the opportunity to give open feedback. It turned out that this kind of system is not suitable for every user and every use case. For example, one of the experts said: “Okay for an overview in an area, but not suitable for the practical work in the orchards”. For other experts, however, it is exactly the kind of tool they have wished for, for example: “Important tool for assessing the increasing influence of parasitoids in the future”. This feedback suggests that our system is a step in the right direction, although it is not suitable for all experts and their use cases. However, some of the experts see the potential benefits of this tool and also an extension to other types of data.

8 Discussion and Future Work

The results of our evaluation make it clear that there is a strong need for intuitive and interactive systems, which support the experts in their daily analysis tasks. The experts are, for the most part, very supportive of *Drosphigator*. Our glyph design was comprehensible, helped them to understand the temporal occurrence of D. suzukii and integrating it in a map helped them to interpret the results. Additionally, allowing for a seamless clustering of vineyards into larger regions is deemed important, as it allows the analysis of micro- and macroecological factors. However, experts are still divided in their opinion, whether the application can support them in their work. This is reflected in their opinion about the possibility to infer causes for the occurrence of D. suzukii from our application.

We plan to improve our system in the future, for example, by integrating a more sophisticated classification algorithm, which considers not only land use and meteorological data, but also different host plant species than just wine. Additionally, we will improve our interaction capabilities and enable experts to integrate their domain knowledge in a Visual Analytics loop, supporting them to better infer causes for the occurrence of D. suzukii.

The ongoing developments of data platforms such as Drosomon, which offer an increasing amount of measurements, with more details and a higher temporal resolution, allow the extension of our approach to support adjustable time-granularities. This enables the analysis of short time periods (days) to investigate acute infestations or long time periods (years) allowing analysis of the effectiveness of taken countermeasures. Additionally, we plan to reduce the occlusion introduced through our glyph by investigating advanced alternative visualization techniques.

For additional future work, we aim to investigate the applicability of our system for spatio-temporal event analysis of other (invasive) species. One particular use case will be the Global Initiative for Honey Bee Health (GIHH) launched by the CSIRO in 2015 [56] aiming to collect scientific evidence of honey bee population decline through global collaboration. Towards this end, microsensors are attached to the bees to record their activity from which predictors of health are inferred. A visual analytics framework [57] is being developed that facilitates interactive analysis of the microsensing data and aids in finding correlates with environmental factors that may impact on bee health. The system we are presenting here is considered a valuable means of visually investigating the health predictors and their related uncertainties on a global scale.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2877352, IEEE Transactions on Big Data

9 Conclusion

In this paper, we presented our application *Drosophigator* which enables researchers in the field of viticulture and biology to investigate the spread dynamics of invasive species. Using data provided by the WBI, we trained an ensemble of classifiers to identify places and times which are susceptible to infestation by D. suzukii. Using our glyph-based visualization, we allow a visual analysis of these spatio-temporal event predictions. We demonstrated the capabilities of our approach in two use-cases, where we show how our tool can be used to investigate hypothesis about the spread of D. suzukii. Additionally, we provide an evaluation of our application by nearly 40 domain experts, which shows the effectiveness of our proposed glyph-based visualization and the visual-interactive system.

Acknowledgments

The project BigGIS (reference number: 01IS14012) is funded by the Federal Ministry of Education and Research (BMBF) within the frame of the research programme “Management and Analysis of Big Data” in “ICT 2020 – Research for Innovations”.

References

[1] A. Cini, C. Ioriatti, G. Anfora *et al.*, “A review of the invasion of drosophila suzukii in europe and a draft research agenda for integrated pest management,” *Bulletin of insectology*, vol. 65, no. 1, pp. 149–160, 2012.

[2] A. Cini, G. Anfora, L. Escudero-Colomar, A. Grassi, U. Santosuosso, G. Seljak, and A. Papini, “Tracking the invasion of the alien fruit pest drosophila suzukii in europe.” *Journal of Pest Science*, vol. 87, no. 4, 2014.

[3] M. K. Asplen, G. Anfora, A. Biondi, D.-S. Choi, D. Chu, K. M. Daane, P. Gibert, A. P. Gutierrez, K. A. Hoelmer, W. D. Hutchison *et al.*, “Invasion biology of spotted wing drosophila (drosophila suzukii): a global perspective and future priorities,” *Journal of Pest Science*, vol. 88, no. 3, pp. 469–494, 2015.

[4] L. Kaiser, “Drosophila suzukii - wissenswertes und lösungsansätze aus der forschung,” 2015.

[5] N. G. Wiman, V. M. Walton, D. T. Dalton, G. Anfora, H. J. Burrack, J. C. Chiu, K. M. Daane, A. Grassi, B. Miller, S. Tochen *et al.*, “Integrating temperature-dependent life table data into a matrix projection model for drosophila suzukii population estimation,” *PLoS One*, vol. 9, no. 9, p. e106909, 2014.

[6] R. Jakobs, T. D. Gariepy, and B. J. Sinclair, “Adult plasticity of cold tolerance in a continental-temperate population of drosophila suzukii,” *Journal of insect physiology*, vol. 79, pp. 1–9, 2015.

[7] M. V. Rossi-Stacconi, R. Kaur, V. Mazzoni, L. Ometto, A. Grassi, A. Gottardello, O. Rota-Stabelli, and G. Anfora, “Multiple lines of evidence for reproductive winter diapause in the invasive pest drosophila suzukii,” *Journal of pest science*, vol. 89, no. 3, pp. 689–700, 2016.

[8] T. Haye, P. Girod, A. Cuthbertson, X. Wang, K. Daane, K. Hoelmer, C. Baroffio, J. Zhang, and N. Desneux, “Current swd ipm tactics and their practical implementation in fruit crops across different regions around the world,” *Journal of pest science*, vol. 89, no. 3, pp. 643–651, 2016.

[9] D. Cormier, J. Veilleux, and A. Firlej, “Exclusion net to control spotted wing drosophila in blueberry fields,” *IOBC-WPRS Bull*, vol. 109, pp. 181–184, 2015.

[10] E. Del Fava, C. Ioriatti, and A. Melegaro, “Cost-benefit analysis of controlling the spotted wing drosophila (drosophila suzukii (matsumura)) spread and infestation of soft fruits in trentino, northern italy,” *Pest Management Science*, 2017.

[11] K. M. Daane, X.-G. Wang, A. Biondi, B. Miller, J. C. Miller, H. Riedl, P. W. Shearer, E. Guerrieri, M. Giorgini, M. Buffington *et al.*, “First exploration of parasitoids of drosophila suzukii,” *Journal of pest science*, vol. 89, no. 3, pp. 823–835, 2016.

[12] X.-G. Wang, G. Kaçar, A. Biondi, and K. M. Daane, “Foraging efficiency and outcomes of interactions of two pupal parasitoids attacking the invasive spotted wing drosophila,” *Biological Control*, vol. 96, pp. 64–71, 2016.

[13] “VitiMeteo,” http://www.vitimeteo.de/, accessed: 12/03/2017.

[14] “DrosoMon,” http://drosophila.julius-kuehn.de/index.php?menuid=44, accessed: 12/03/2017.

[15] D. A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler, “Challenges in visual data analysis,” in *Information Visualization, 2006. IV 2006. Tenth International Conference on*. IEEE, 2006, pp. 9–16.

[16] N. Andrienko, G. Andrienko, and P. Gatalsky, “Exploratory spatio-temporal visualization: an analytical review,” *Journal of Visual Languages & Computing*, vol. 14, no. 6, pp. 503–541, 2003.

[17] S. McKenna, D. Staheli, C. Fulcher, and M. Meyer, “Bubblenet: A cyber security dashboard for visualizing patterns,” in *Computer Graphics Forum*, vol. 35, no. 3. Wiley Online Library, 2016, pp. 281–290.

[18] A. Thudt, D. Baur, and S. Carpendale, “Visits: A spatiotemporal visualization of location histories,” in *Proceedings of the eurographics conference on visualization*, 2013.

[19] A. Mehler, Y. Bao, X. Li, Y. Wang, and S. Skiena, “Spatial analysis of news sources,” *IEEE transactions on visualization and computer graphics*, vol. 12, no. 5, pp. 765–772, 2006.

[20] R. Maciejewski, S. Rudolph, R. Hafen, A. Abusalah, M. Yakout, M. Ouzzani, W. S. Cleveland, S. J. Grannis, and D. S. Ebert, “A visual analytics approach to understanding spatiotemporal hotspots,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 2, pp. 205–220, 2010.

[21] B. N. Hilton, T. A. Horan, R. Burkhard, and B. Schooley, “Saferoadmaps: Communication of location and density of traffic fatalities through spatial visualization and heat map analysis,” *Information Visualization*, vol. 10, no. 1, pp. 82–96, 2011.

[22] R. Maciejewski, R. Hafen, S. Rudolph, W. S. Cleveland, and D. S. Ebert, “Forecasting hotspots-a predictive visual analytics approach.”

[23] M. Stein, H. Janetzko, A. Lamprecht, D. Seebacher, T. Schreck, D. A. Keim, and M. Grossniklaus, “From Game Events to Team Tactics: Visual Analysis of Dangerous Situations in Multi-Match Data,” in *2016 1st International Conference on Technology and Innovation in Sports, Health and Wellbeing (TISHW)*, Dec. 2016, pp. 1–9, special Track ”High level Sports in the XXI Century: contribution from Industry and University to the performance optimization”.

[24] C. Rohrdantz, M. Krstajic, M. El Assady, and D. Keim, “What’s going on? how twitter and online news can work in synergy to increase situational awareness,” 2012.

[25] J. Fuchs, F. Fischer, F. Mansmann, E. Bertini, and P. Isenberg, “Evaluation of alternative glyph designs for time series data in a small multiple setting,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI ’13. New York, NY, USA: ACM, 2013, pp. 3237–3246.

[26] D. A. Keim, “Designing pixel-oriented visualization techniques: Theory and applications,” *IEEE Transactions on visualization and computer graphics*, vol. 6, no. 1, pp. 59–78, 2000.

[27] M. Weber, M. Alexa, and W. Müller, “Visualizing time-series on spirals.” in *Infovis*, vol. 1, 2001, pp. 7–14.

[28] J. V. Carlis and J. A. Konstan, “Interactive visualization of serial periodic data,” in *Proceedings of the 11th annual ACM symposium on User interface software and technology*. ACM, 1998, pp. 29–38.

[29] A. Diehl, L. Pelorosso, C. Delrieux, C. Saulo, J. Ruiz, M. E. Gröller, and S. Bruckner, “Visual analysis of spatio-temporal data: Applications in weather forecasting,” in *Computer Graphics Forum*, vol. 34, no. 3. Wiley Online Library, 2015, pp. 381–390.

[30] G. Fuchs and H. Schumann, “Visualizing abstract data on maps,” in *Information Visualisation, 2004. IV 2004. Proceedings. Eighth International Conference on*. IEEE, 2004, pp. 139–144.

[31] J. Zhao, P. Forer, and A. S. Harvey, “Activities, ringmaps and geovisualization of large human movement fields,” *Information visualization*, vol. 7, no. 3-4, pp. 198–209, 2008.

[32] “Biologie der Drosophila Suzukii,” https://drosophila.julius-kuehn.de/index.php?menuid=28, accessed: 2018-06-06.

[33] E. Pelton, C. Gratton, R. Isaacs, S. Van Timmeren, A. Blanton, and C. Guédot, “Earlier activity of drosophila suzukii in high woodland landscapes but relative abundance is unaffected,” *Journal of Pest Science*, vol. 89, no. 3, pp. 725–733, Jul 2016. [Online]. Available: http://dx.doi.org/10.1007/s10340-016-0733-z

[34] “ASTER Global Digital Elevation Map,” https://asterweb.jpl.nasa.gov/gdem.asp, accessed: 2018-06-06.

[35] D. Grünreich, “Atkis-a topographic information system as a basis for gis and digital cartography in germany,” *From Digital Map Series to Geo-Information Systems, Geologisches Jarhrbuch Series A. Hannover, Germany: Federal Institute of Geosciences and Resources*, 1992.

[36] “ALKIS-Objektartenkatalog (2008),” https://www.lgl-bw.de/lgl-internet/web/sites/default/de/07_Produkte_und_Dienstleistungen/Galerien/Dokumente/ALKIS_OK_6_0.pdf, accessed: 2018-06-06.

[37] W. Liu, S. Chawla, D. A. Cieslak, and N. V. Chawla, "A robust decision tree algorithm for imbalanced data sets," in *Proceedings of the 2010 SIAM International Conference on Data Mining*. SIAM, 2010, pp. 766–777.

[38] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1, pp. 1–39, 2010.

[39] M. Skeels, B. Lee, G. Smith, and G. G. Robertson, "Revealing uncertainty for information visualization," *Information Visualization*, vol. 9, no. 1, pp. 70–81, 2010.

[40] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[41] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel, "KNIME: The Konstanz Information Miner," in *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer, 2007.

[42] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[43] D. G. Altman, *Practical statistics for medical research*. CRC press, 1990.

[44] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

[45] T. K. Ho, J. J. Hull, and S. N. Srihari, "Decision combination in multiple classifier systems," *IEEE transactions on pattern analysis and machine intelligence*, vol. 16, no. 1, pp. 66–75, 1994.

[46] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[47] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Visual Languages, 1996. Proceedings., IEEE Symposium on*. IEEE, 1996, pp. 336–343.

[48] N. Ferreira, L. Lins, D. Fink, S. Kelling, C. Wood, J. Freire, and C. Silva, "Birdvis: Visualizing and understanding bird populations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2374–2383, 2011.

[49] R. Borgo, J. Kehrer, D. H. Chung, E. Maguire, R. S. Laramee, H. Hauser, M. Ward, and M. Chen, "Glyph-based visualization: Foundations, design guidelines, techniques and applications." in *Eurographics (STARs)*, 2013, pp. 39–63.

[50] K. Moreland, "Diverging color maps for scientific visualization," *Advances in Visual Computing*, pp. 92–103, 2009.

[51] "Leaflet - Markercluster," https://github.com/Leaflet/Leaflet.markercluster, accessed: 2017-07-06.

[52] L. A. dos Santos, M. F. Mendes, A. P. Krüger, M. L. Blauth, M. S. Gottschalk, and F. R. Garcia, "Global potential distribution of drosophila suzukii (diptera, drosophilidae)," *PloS one*, vol. 12, no. 3, p. e0174318, 2017.

[53] "6th workshop of the working group "D. Suzukii'," http://www.zepp.info/termine-a-news/219-6-arbeitstreffen-der-ag-kirschessigfliege, accessed: 2018-06-06.

[54] R. Likert, "A technique for the measurement of attitudes." *Archives of psychology*, 1932.

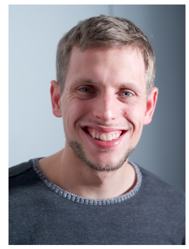
[55] R. M. Heiberger, N. B. Robbins *et al.*, "Design of diverging stacked bar charts for likert scales and other applications," *Journal of Statistical Software*, vol. 57, no. 5, pp. 1–32, 2014.

[56] "Global Initiative for Honey Bee Health (GIHH)," https://research.csiro.au/gihh/about/, accessed: 07/03/2017.

[57] U. Engelke, P. Marendy, F. Susanto, R. Williams, S. Mahbub, H. Nguyen, and P. de Souza, "A visual analytics framework to study honey bee behaviour," in *Proceedings of the 2nd IEEE Data Science and Systems Conference (DSS)*. IEEE, 2016, pp. 1504–1511.

Daniel Seebacher is a research associate and Ph.D. student at the Data Analysis and Visualization group of Daniel Keim at the University of Konstanz since 2016. His research interests include similarity search in heterogeneous high-dimensional data. He received an MSc. in Computer Science from the University of Konstanz in 2015.

Johannes Häußler is a research associate and Ph.D. candidate at the Data Analysis and Visualization group of Daniel Keim at the University of Konstanz since 2017. His research interests include geographic information systems and the visualization of geo-temporal phenomena. He received his M.Sc. in Computer Science from the University of Konstanz.

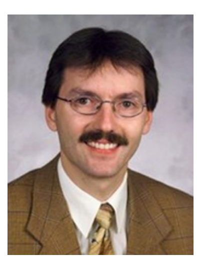


Michael Hundt is a master student in the Data Analysis and Visualization Group at the University of Konstanz. His research interests include computational social science and spatiotemporal text data analysis, in particular pattern detection and visualization of social media. Hundt has a BS in computer science from the University of Konstanz.

Manuel Stein is a PhD student in the Data Analysis and Visualization Group at the University of Konstanz. His research interests include movement analysis and team sports analytics, in particular pattern detection, decision-making processes, and context-aware visual analysis. Stein has an MS in computer science from the University of Konstanz.

Hannes Müller is a scientific employee at the State Institute for Environment of the Federal State of Baden-Württemberg (LUBW) since 2016. His activities include management of IT-projects for the administration and distribution of environmental data. He received a PhD in Geography from the Humboldt-Universität zu Berlin.

Ulrich Engelke is a Senior Research Scientist with DATA61 at the Commonwealth Scientific and Industrial Research Organization (CSIRO) and an Adjunct Senior Lecturer at the University of Tasmania, Australia. He is co-founder and steering committee member of the International Symposium for Big Data Visual Analytics (BDVA) and a Senior Member of the IEEE. His research interests include visual analytics, information visualisation, and immersive user experiences.



Daniel A. Keim is full professor and head of the Data Analysis and Visualization Research Group at the University of Konstanz's Computer Science Department. His research combines automated data analysis and interactive visualization techniques for the purpose of exploring large data in various applications domains. Daniel Keim has a PhD in computer science from the University of Munich