Crowdsourcing Insect Observation to Assess Demographic Shifts and Improve Classification

Léonard Boussioux [1 2] **Tomás Giro-Larraz** [2 3] **Charles Guille-Escuret** [1] **Mehdi Cherti** [4 5 6] **Balázs Kégl** [4 5]

Abstract

Insects play such a crucial role in ecosystems that a shift in demography of just a few species can have devastating consequences at environmental, social and economic levels. Despite this, evaluation of insect demography is strongly limited by the difficulty of collecting census data at sufficient scale. We propose a method to gather and leverage observations from bystanders, hikers, and entomology enthusiasts in order to provide researchers with data that could significantly help anticipate and identify environmental threats. Finally we show that there is indeed interest on both sides for such collaboration.

1. Introduction

It is estimated that 90% of all animal life forms on Earth are insects (Erwin, 1982; 1997) with a ratio of 200 millions specimens per human being (Pedigo & Rice, 2009) at any time, and that there exists 6 to 10 million different species (Chapman, 2006), of which we have only identified 900,000. With such numbers, it is tremendously difficult for entomologists to classify insect species, estimate current population distribution and detect their shifts. Yet, habitat loss, intensification of agricultural practices, urbanization and environmental changes result in insect decline (Dupont & Olesen., 2009; Vanbergen, 2013) and demographic shifts with dramatic consequences:

The Asian Hornet (*Vespa velutina*) is an invasive species (Tan et al., 2007) with a population quickly increasing in Europe and particularly in France. It was introduced to this new habitat by humans and is a major concern due to several hospitalizations related to their stings and their aggressiveness toward local fauna, effectively destabilizing local ecosystems

The Tiger Mosquito (*Aedes albopictus*) is another widespread invasive species in the world. Its outstanding adaptation ability being boosted by global warming, it has outcompeted concurrent species on several continents. It is a vector for many dangerous diseases (yellow fever, dengue fever, Chikungunya fever and Usutu virus), making it a serious health concern worldwide.

Figure 1. Asian hornet, tiger mosquito and honey bee.

Honey bees (Genus *Apis*) have been disappearing at alarming rates in the first decade of the 21st century, due to a phenomenon called Colony Collapse Disorder. The causes are still unclear, but are suspected to be a mix of environmental perturbations (vanEngelsdorp et al., 2009). Because agriculture depends heavily on bees and other pollinators, monitoring these populations is critical to anticipating and preventing the disasters that would follow large population declines.

Besides the direct impact of some demographic shifts, insect population distributions can also be used as powerful indicators of the ecosystem perturbations such as global warming, destruction of habitat, pesticides and pollution to justify environment protection policies and measure their effectiveness (Renato Mauricio da Rocha et al., 2011).
Since the scale of the problem makes it intractable for clas-

[1]Montreal Institute for Learning Algorithms, Montréal, Canada [2]Ecole CentraleSupélec, Gif-sur-Yvette, France [3]Ecole Polytechnique Fédérale de Lausanne, Lausanne, Suisse [4]CNRS, Orsay, France [5]Université Paris-Saclay, Orsay, France [6]Mines ParisTech, Paris, France. Correspondence to: Léonard Boussioux <leobix@mit.edu>, Charles Guille-Escuret <charles.guille-escuret@umontreal.ca>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, USA, 2019.

sical census methods, we propose an approach based on volunteer contributions from amateurs and bystanders, who benefit in return from insect identification tools, playful features and the appeal of contributing to crucial statistical studies.

2. Related Work

We identify in particular two successful similar approaches relying on crowd-sourcing observations of living beings encouraged by an identification tool : Pl@ntnet, a mobile application for plant observation and identification, and iNaturalist, a platform encompassing all living beings. With roughly 500,000 users having made more than 18,000,000 observations over 210,000 species, it is a very popular tool that has proven public interest in this area.
However, the extremely large size of iNaturalist's scope has some drawbacks when dealing with insects specifically. Pictures that include both insects and plants may be difficult to identify due to their ambiguous nature, and because of how small the differences between species can be, insect manual identification sometimes requires specific expertise. Moreover, entomologists have expressed their strong interest for observation protocols specifically designed for statistical studies of insects (see section 3.4), which can be greatly facilitated by a dedicated tool.

3. Method

3.1. The Mobile Application *InsectUp*

Similarly to the *Galaxy Zoo project* (Lintott et al., 2010), which proposed astronomy amateurs to help categorizing galaxies and resulted in 40 million handmade classifications within 6 months, we seek to leverage the contributions made by users to help entomologists in their large scale tasks.

This is made through a mobile phone application called *InsectUp*, which offers to identify insects on pictures taken by users. It also provides educational information on the identified insects and local fauna. The identified pictures can be used in turn as observations, to be used as demographic data along with picture location.
Additionally, the application offers to enthusiasts the possibility to help identifying insects on newly collected pictures. Along with a verification process to limit erroneous annotations, it allows a continual enrichment of the dataset with new samples and species, effectively improving its own classification capabilities.

3.2. Classification Data

The initial dataset is provided by the French Photographic Survey of Flower Visitors (SPIPOLL) (de Flores & Deguines, 2012), a project sponsored by the French National Museum of Natural History and Office for Insects and their Environment, of which a few examples are presented in Figure 2. It contains 145k crowd-sourced labelled observations over 403 species of insects. While the number of species is low, it is due to the combination of biological and geographical constraints : it only considers pollinating insects in mainland France. As more data is collected and manually annotated, the model can be retrained to increase its number of recognized species and accuracy.

Figure 2. Example of SPIPOLL pictures obtained by capturing all visitors of an iris flower during 20 minutes.

3.3. Classification Model

We rely on the RAMP framework (Kégl et al., 2018), a platform developed for building transparent pipelines and easily reproducible results, and use as a model Inception-v4 (Szegedy et al., 2015; 2016), pretrained on ImageNet (Deng et al., 2009) then fine-tuned on our dataset. For training, we used a batch size of 32 and trained the model for 8 epochs. Each image was first center-cropped then resized to 224×224. We used standard SGD with an initial learning rate of 0.01. We divide the learning rate by 10 in the 4th epoch and the 7th epoch. To avoid overfitting, we used several data augmentation techniques such as horizontal and vertical flipping, random perturbations of contrast, random shift, random scale and random rotation.
We obtain an 87% top-1 accuracy on the SPIPOLL test set ($\sim$ 70k images).

3.4. Interest for the Research Community

The SPIPOLL project was originally launched as an effort from French entomologists to encourage citizen scientists to gather data on pollinating insects, following a standardized protocol, to assess macroecological changes in richness and composition of flower visitor communities. This protocol will typically involve taking pictures of all insect visitors of a flower for a certain duration, then manually upload, identify and annotate each of these pictures to the SPIPOLL website. Such procedure can be greatly facilitated by the automation of both upload and identification of the pictures, and improving insect identification performance could lead to a full automation of the process through autonomous cameras, which would multiply current insect census capabilities. The French National Center for Scientific Research (CNRS) and the French National Museum of Natural History have expressed their strong interest for data collected through such protocols.
Furthermore, even without observation protocols or autonomous cameras, the generated data can help estimate some demographic metrics such as the geographic distribution of a given species or its fluctuation over time. It is not implausible either to make original observations of species in regions were they had not been observed before or in time periods when they were not known as present.
However, it is important to acknowledge that this data may not be used directly to numerically estimate populations: indeed, it is subject to a strong observer bias. Species that are more commonly seen in homes and urban areas, that are easy to notice because of their size and color, or are easier to capture (e.g. because they don't fly) will, for example, be over-represented in the collected data.
Finally, we have partnered with the Paris-Saclay Center for Data Science which provides us with technical support, access to their servers, and Amazon AWS funding.

4. Results

4.1. Interest from bystanders and entomology amateurs

The mobile application *InsectUp* was launched in April 2018 on the Android Playstore only, without any marketing. This launch served as an alpha to test its features and interest from users.
Despite the lack of visibility and the very limited features proposed by the app at this time, it had an unexpected success with over 50,000 downloads and more than 8,000 monthly active users at its peak in the middle of the Summer (Figure 3). Over the 7 months of the test, a total of more than 44,000 insect pictures have been uploaded by users (Figure 4), which shows that people are curious to identify insects and ready to participate actively. Additionally there were many comments, identifications and discussions as the community progressively developed.
This proves the potential of this tool to reach a significant number of users, an essential condition to obtain data of sufficient scale.

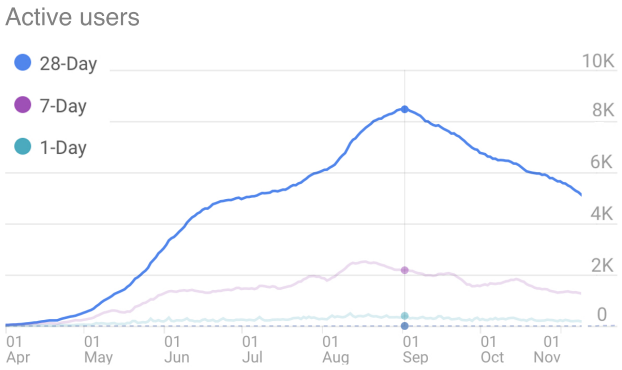

Figure 3. 1-day, 7-day, and 28-day active users from April 2018 launch to November 2018, when the application was removed from the Playstore for refactoring.

Figure 4. 15 pictures taken by the users of our app and posted to our public gallery. Photo quality and insect species are very variable.

4.2. User profile

An interesting takeaway is that this project attracted a very wide community. Figure 5 shows that our application was downloaded all around the world, by users of all ages. It did not appeal only to entomology amateurs, but also to curious bystanders.

5. Discussion

While we believe to have shown the strong potential of this application for social good, a few technical challenges will need to be solved to fully take advantage of crowd-

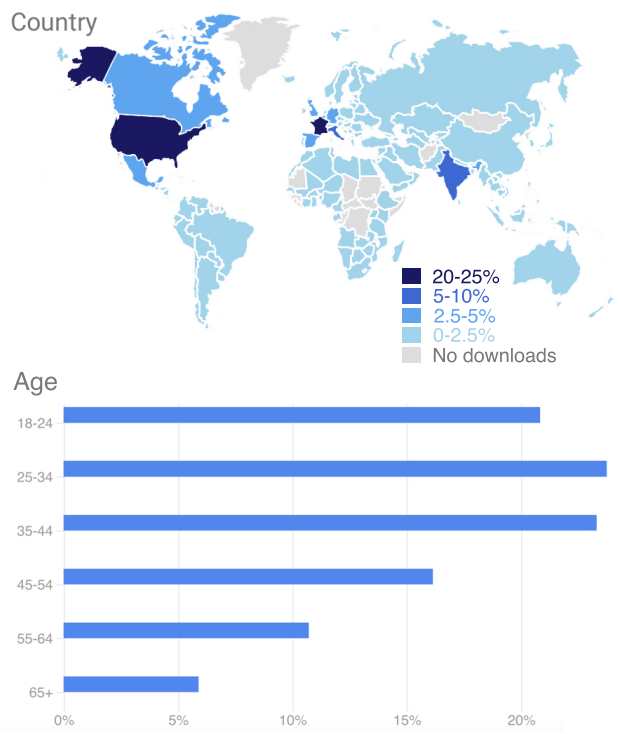

Figure 5. Age and geographic distribution of the application users.

sourced observations. We present here 3 of these challenges.

With nearly a million insect species known on Earth, it is to be expected that the amount of data available per species will be highly variable. Most species have never been classified by entomologists, and others are overwhelmingly represented due to their high population and easy observability. As the dataset grows in diversity, many species will only have a few annotated samples within the dataset, which can be extremely challenging to tackle, especially on such a very large number of classes. The ability of the algorithm to quickly pick up on these new species to enrich the variety of its identifications is critical. While few-shot learning has recently made great progress (Vinyals et al., 2016; Finn et al., 2017), these advances are largely targeted at *n-shot*, *k-way* meta-learning tasks, where the support dataset is part of the input. The size of our training data and the high variance in sample size between species call for a different solution. For example, the ability to perform a less refined classification for underrepresented species, such as identifying the genus or the family, could be an acceptable solution.

Since our approach is largely built around manual annotations from humans with different levels of expertise, a rigorous annotation pipeline needs to be designed to avoid erroneous identifications. This is especially difficult in the insect classification case because of the high level of similarity between some species. A good estimation of annotation confidence based on multiple identification suggestions and users identification history seems necessary, with possibly the intervention of a trained entomologist for disputed observations. While ensuring a satisfying reliability of the labels at a high scale may be challenging, experience has shown there is an active community of entomologists willing to perform high quality identifications if the interface is well designed.

Another issue is false observations, which can for example be produced by users taking pictures of themselves to see which insect they supposedly look like, as has happened a couple of times during the alpha phase. Additionally, a simple way to increase users engagement would be to introduce game-like features, e.g. encouraging to take pictures of as many different local species as possible. However this kind of feature would push some users to upload pictures found on the web, effectively degrading data quality. As a result, a protection against false observations would be greatly beneficial.

While there are other directions to explore that would improve the potential of this application, we believe these 3 are some of the most critical for the large scale success of our approach.

6. Conclusion

We have presented the early development of a crowdsourced insect observation tool, with the aim of providing entomologists high scale demographic data on a wide range of insects. We have shown evidence of the interest from both the scientific community and amateurs to collaborate in a platform that could improve our understanding and assessment of insect populations, and shown how such knowledge largely benefits society through its critical role in environmental protection.
While there is still a lot to do on this specific application, we believe similar approaches based on collaboration between citizen scientists and researchers could be successfully applied to many fields where gathering data is expensive and critical.

Acknowledgements

We thanks Colin Fontaine, Grégoire Loïs, Romain Julliard and the French National Museum of Natural History for sharing the dataset and valuable insights on their needs. We also thanks Rachel Tourneix, Mathilde Bryant, Jason

Boussioux, Martin Kégl for their precious feedback on the mobile application and Louis Maestrati, Anne-Flore Baron, Baptiste Goujaud for their help in reviewing the paper. This work was supported by a grant from the *Associations des Centraliens et des Supélec of Languedoc-Roussillon* and from *La Recherche* and *Sciences et Avenir*.

References

Chapman, A. *Numbers of living species in Australia and the World*. 2nd edition, 2006.

de Flores, M. and Deguines, N. Trois ans d'activité du spipoll. *Insectes*, 167:9–12, 12 2012.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Dupont, Y. L. and Olesen., J. M. Ecological modules and roles of species in heathland plant-insect flower visitor networks. *Journal of Animal Ecology*, 78:346–353, 2009.

Erwin, T. L. Tropical forests: their richness in coleoptera and other arthropod species. *The Coleopterists Bulletin*, 36:74–75, 1982.

Erwin, T. L. Biodiversity at its utmost: Tropical forest beetles. In Reaka-Kudla, M.L.; Wilson, D. W. E. (ed.), *Biodiversity II.*, pp. 27–40. Joseph Henry Press, Washington, D.C., 1997.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. *CoRR*, abs/1703.03400, 2017. URL http://arxiv.org/abs/1703.03400.

Kégl, B., Boucaud, A., Cherti, M., Kazakci, A., Gramfort, A., Lemaitre, G., Van den Bossche, J., Benbouzid, D., and Marini, C. The RAMP framework: from reproducibility to transparency in the design and optimization of scientific workflows. *ICML*, 2018.

Lintott, C., Schawinski, K., Bamford, S., Slosar, A., Land, K., Thomas, D., Edmondson, E., Masters, K., Nichol, R., Raddick, J., Szalay, A., Andreescu, D., Murray, P., and Vandenberg, J. Galaxy zoo 1 : Data release of morphological classifications for nearly 900,000 galaxies. *Monthly Notices of the Royal Astronomical Society*, 2010.

Pedigo, L. and Rice, M. *Entomology and Pest Management*. Pearson Prentice Hall, 6th edition, 2009.

Renato Mauricio da Rocha, J., Almeida, J., Lins, G., and Durval, A. Insects as indicators of environmental changing and pollution: A review of appropriate species and their monitoring. *Holos Environment*, 10, 07 2011. doi: 10.14295/holos.v10i2.2996.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich., A. Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. *CVPR*, 2016.

Tan, K., Radloff, S. E., Li, J. J., Hepburn, H. R., Yang, M. X., Zhang, L. J., and Neumann, P. Bee-hawking by the wasp, vespa velutina, on the honeybees apis cerana and a. mellifera. *Naturwissenschaften*, 94, 2007.

Vanbergen, A. J. Threats to an ecosystem service: pressures on pollinators. *Frontiers in Ecology and the Environment*, 11:251–259, 2013.

vanEngelsdorp, D., Evans, J. D., Saegerman, C., Mullin, C., Haubruge, E., Nguyen, B. K., Frazier, M., Frazier, J., Cox-Foster, D., Chen, Y., Underwood, R., Tarpy, D. R., and Pettis, J. S. Colony collapse disorder: A descriptive study. *PLOS ONE*, 4(8):1–17, 08 2009. doi: 10.1371/journal.pone.0006481. URL https://doi.org/10.1371/journal.pone.0006481.

Vinyals, O., Blundell, C., Lillicrap, T. P., Kavukcuoglu, K., and Wierstra, D. Matching networks for one shot learning. *CoRR*, abs/1606.04080, 2016. URL http://arxiv.org/abs/1606.04080.