ERDC/GRL TN-20-5
September 2020

Understanding the Disease Vector Operational Environment by Predicting Presence of *Anopheles* Mosquito Breeding Sites Using Maximum Entropy Modeling and the Maxent Software Platform

By Susan L. Lyon and Kathleen V. Payne

PURPOSE: This technical note (TN) describes research using the maximum entropy model to predict the presence of breeding sites for mosquitos of the genus *Anopheles* throughout the Korean peninsula. This methodology is also applicable to many other types of ecological niche modeling problems where analysts only have access to data related to the location a species has been found.

The purpose of this study is to help address the need for new and innovative methods that promote military readiness through better understanding of vector-borne disease threats in familiar and unfamiliar operational environments. These methods can be used to provide military planners with valuable information to support their operations, particularly when operations expand into areas lacking direct disease vector surveillance. Disease vector risk information is vital for force readiness, because historically, soldiers are more likely to be unable to perform warfighting due to disease and non-combat injuries than as a direct result of combat (U.S. Department of the Army 2015).

INTRODUCTION: The *Anopheles* genus is comprised of several hundred species of mosquito, dozens of which have the ability to transmit the parasites that cause malaria in humans (Kim et al. 2011). In 2017, there were an estimated 219 million cases of malaria worldwide, and an estimated 435,000 deaths from the disease that same year (World Health Organization 2019). Malaria is of particular interest to Army medical planners because of the long history of adverse effects it has had on combat operations (Kim et al. 2011; U.S. Department of the Army 2015). Military doctors diagnosed and treated an estimated 12,000 cases of malaria among U.S. military personnel in Korea in the early 1950s (Fukuda et al. 2018). In 1993, the disease reemerged in South Korea after a two-decade absence due to a long-term government eradication program (Cho et al. 1994). Since then, the Army has continued surveillance and study of *Anopheles* in Korea, along with assessing malaria risk to force readiness and evaluating prevention methods (Klein et al. 2008).

While malaria in South Korea is relatively rare in the modern era, the models developed in this study provide a glimpse into the potential presence of *Anopheles* breeding sites in other areas that may not have extensive malaria eradication programs or surveillance data, such as North Korea or China.

Vector control is the primary method used to lower the risk of malaria (World Health Organization 2019). Because some vector control methods work better at certain life stages, effective planning and use of such controls relies on an understanding of the life cycle and habitat of *Anopheles* mosquito species in the area of interest. The general life cycle of mosquitos, which includes species under the genus *Anopheles*, is illustrated in Figure 1. After eggs have been laid, they develop through four larva stages, shedding the outer skin during each phase. Since some common vector control methods target egg sites, this study focused on modeling the locations of field collections for the first and second larval stages, because the larva are young enough that they have not been moved from the original breeding location by outside forces like moving water.

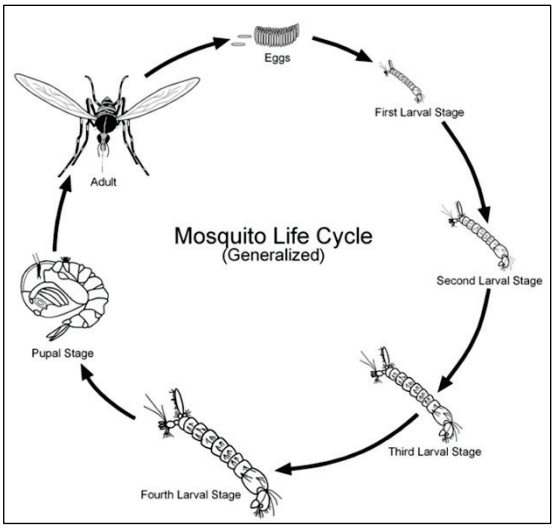

Figure 1. Life cycle of mosquitos (Charlesworth 2008).

METHODS AND DATA

Maximum entropy modeling software. Maximum entropy is a machine learning method that makes predictions based on incomplete information (Phillips et al. 2006). Since the maximum entropy method is not usually implemented in standard statistical software platforms, a dedicated software application called Maxent can be used for analysis. Maxent is a Java-based open-source Windows application that was developed to provide a way to use the maximum entropy method to model species geographic distributions (Phillips et al. 2004). The software uses data from known species presence locations, along with environmental data that describe aspects of the habitat that may cause the species to prefer those locations, and builds a probability map predicting the likelihood the target species will be found throughout the area of interest (Phillips et al. 2006). Maxent can be downloaded at *http://biodiversityinformatics.amnh.org/open_source/maxent/*.

Maxent requires two types of data inputs. First is a presence dataset, recording areas where the subject of the analysis has been found. While other analysis methods may integrate absence data, Maxent only uses presence information. This dataset must be a comma separated value text file (CSV), in which each line contains one species or phenomena that is to be predicted and x and y geographic coordinates (Figure 2). The first column is the species or phenomena being studied; more than one can be used in the analysis, in which case each line would contain one species name, one x coordinate observation, and one y coordinate observation.

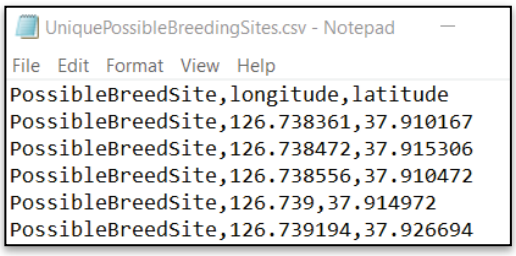

Figure 2. Presence dataset, properly formatted to be used in Maxent software.

The second required data input is one or more environmental datasets that describe the habitat of the species. These must be in the form of rasters, where environmental data are presented as a

value in each cell of a grid. If multiple variables will be used in the analysis, the rasters must exactly match each other in geographic extent and cell size. The geographic coordinate system used in the rasters must match that of the presence dataset. No null values can be contained in the rasters and the dataset must cover all cells within a bounding box; it must also be saved in the ASCII (.asc) format. The specific variables chosen should relate to an aspect of the habitat preferences of the species and may be continuous or categorical variables. Alignment and processing of the raster datasets can be done through many geographic information system (GIS) or data analysis software platforms. This study used ArcGIS's Raster Toolbox to ensure the rasters' geographic extents matched.

Larva data inputs. The presence data used in this study were derived from field collections of *Anopheles* larva and larva shells at multiple sites around Warrior Base in Gyeonggi, Republic of Korea. Dr. Terry Klein and Dr. Heung-Chul Kim, public health entomologists with Force Health Protection and Preventive Medicine, MEDDAC-Korea/65th Medical Brigade, supervised and conducted this data collection from May to October 2010. In addition to manual counts of larva and larval shells found in the field at each collection site, detailed descriptions of the site environments were recorded and photographs taken (Figure 3).

Figure 3. Example larva collection sites in northern South Korea.

Field data were recorded as forms collected in Excel spreadsheets, but were not in a format that was easily translated into data tables for analysis. Therefore, the first step of this study was manually taking these several hundred collection forms and reformating them into a single data table. While this study was only concerned with the locations where larva indications were found, all of the variables were tabulated together, for ease of use in potential future studies.

The resulting data table was then cleaned using the open-source tool OpenRefine, available at *http://openrefine.org/*. OpenRefine allows easy cleaning and standardization of big data sets. In this data set, many descriptions of the collection sites were slightly different because they were entered by field researchers in different ways. These similar values were converted into a single representative value with OpenRefine. The data set was then reduced to only those locations with first or second stage larva, leaving only potential breeding sites. These 28 sites are shown in Figure 4, but with a great deal of overlap due to the close proximity of most of the collection sites. The CSV was then reviewed to verify correct formatting, as seen in Figure 2.

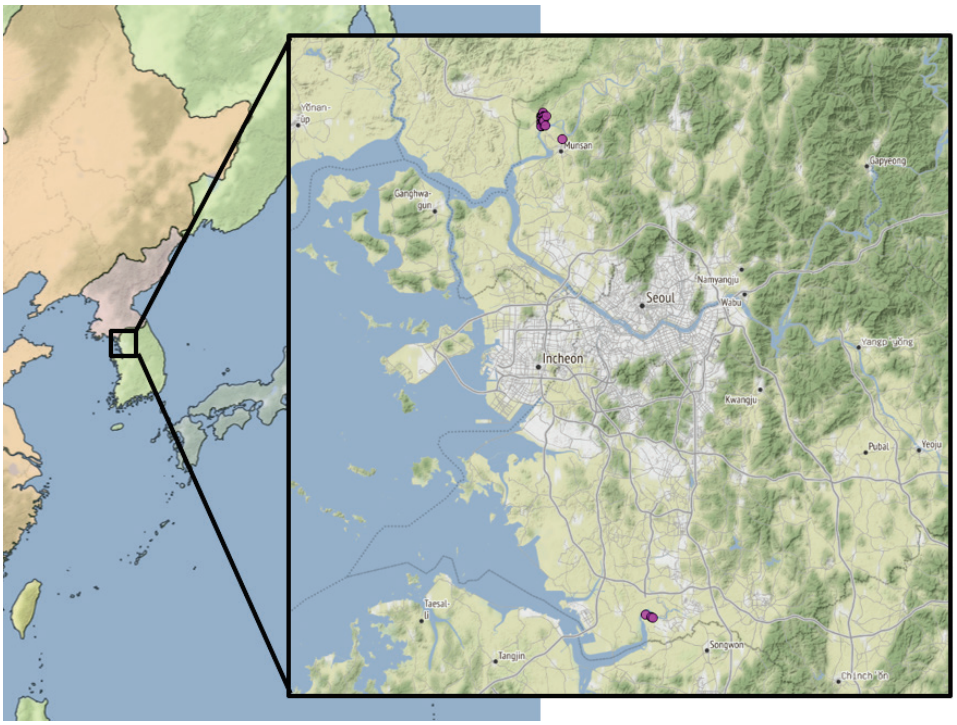

Figure 4. Potential breeding sites of *Anopheles* mosquitos shown in purple.

Environmental data inputs. The Maxent software requires input rasters that describe the habitat suitability of the target species. This study used rasters at 300 m resolution, which was chosen because of the desire for a relatively fine granularity in the output since it is intended to inform tactical-level military medical planning.

The following datasets were used because of the known influence these variables have on the presence of *Anopheles* mosquitos (Kim et al. 2011):

- Mean normalized difference vegetation index (NDVI) as calculated from Landsat 7, averaged over data observed from April 1 through October 31, 2010, courtesy of the U.S. Geological Survey Earth Resources Observation and Science (EROS) Center. NDVI is a measure of the health of vegetation (Rouse et al. 1974).

- Elevation, from the Shuttle Radar Topography Mission (SRTM) (Farr et al. 2007).

- Slope, as calculated from the SRTM dataset.

- Precipitation, in the format of the mean of Climate Hazards Infrared Precipitation with Stations (CHIRPS) data from April 1 2010 through October 31, 2010 (Funk et al. 2015).

- Population count from Worldpop (*http://www.worldpop.org*). This dataset estimates the human population count within 100 m grid squares. (Gaughan et al. 2013; Center for International Earth Science Information Network 2016)

These datasets were downloaded via Google Earth Engine (GEE), which is available at *https://code.earthengine.google.com/*. GEE is an online platform for analysis and visualization of geospatial data and hosts a catalog of geospatial datasets (Gorelick et al. 2017). The rasters were downloaded at a resolution of 300 m after being resampled from their native resolutions, and were clipped to the same bounding box polygon. This bounding box and the cell size of 300 m will also be the dimensions of the output prediction map.

Maxent was then run using the file with the presence point data and the environmental variable rasters. The file paths to the presence data file and the directory containing the environmental data rasters can be entered on the Maxent main screen, shown in Figure 5. Additional options can be found in the Settings screen. A full discussion of the details for the Maxent software is beyond the scope of this TN, but can be found in the Maxent documentation and other previous work (Merow et al. 2013; Griffin 2017).

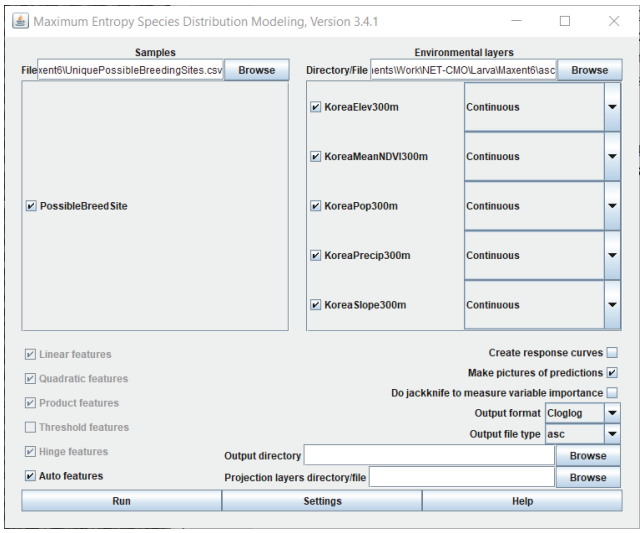

Figure 5. Maxent software main screen.

RESULTS: The output of the Maxent software is shown in Figure 6. Cells with a higher predicted likelihood of the presence of *Anopheles* breeding sites are shown in red, and cells with the lowest likelihood are shown in blue. The presence locations used to build the model are indicated by white squares on the prediction map. The size of the presence squares does not reflect the cell size and is oversized for visibility. While there were 28 presence points in the input dataset, only 19 were used to create the prediction. This is because each cell only counts as one presence record, even where multiple presence points are located within the area of the cell.

There is no specific measure evaluating the maximum entropy model performance within the Maxent software output. However, the specificity curve in Figure 7 gives an indicator of how well the model is working (see Radosavljevic and Anderson [2014] for further discussion of evaluation of Maxent models). The area under the curve (AUC) of a random prediction is 0.5, and a theoretical perfect prediction would be around 1. The AUC of the specificity curve of the Maxent model produced in this study was 0.995, which suggests that the model has performed well. This model evaluation method is not as refined compared to a confidence interval or p value that can be found in conventional statistical models, but it gives some indication that the result is far from random chance. The Maxent software allows the option of additional validation through splitting the presence points into training and testing sets. This study did not use the train-test validation method because there were only 19 presence points; a typical train/test split would have removed too many of them.

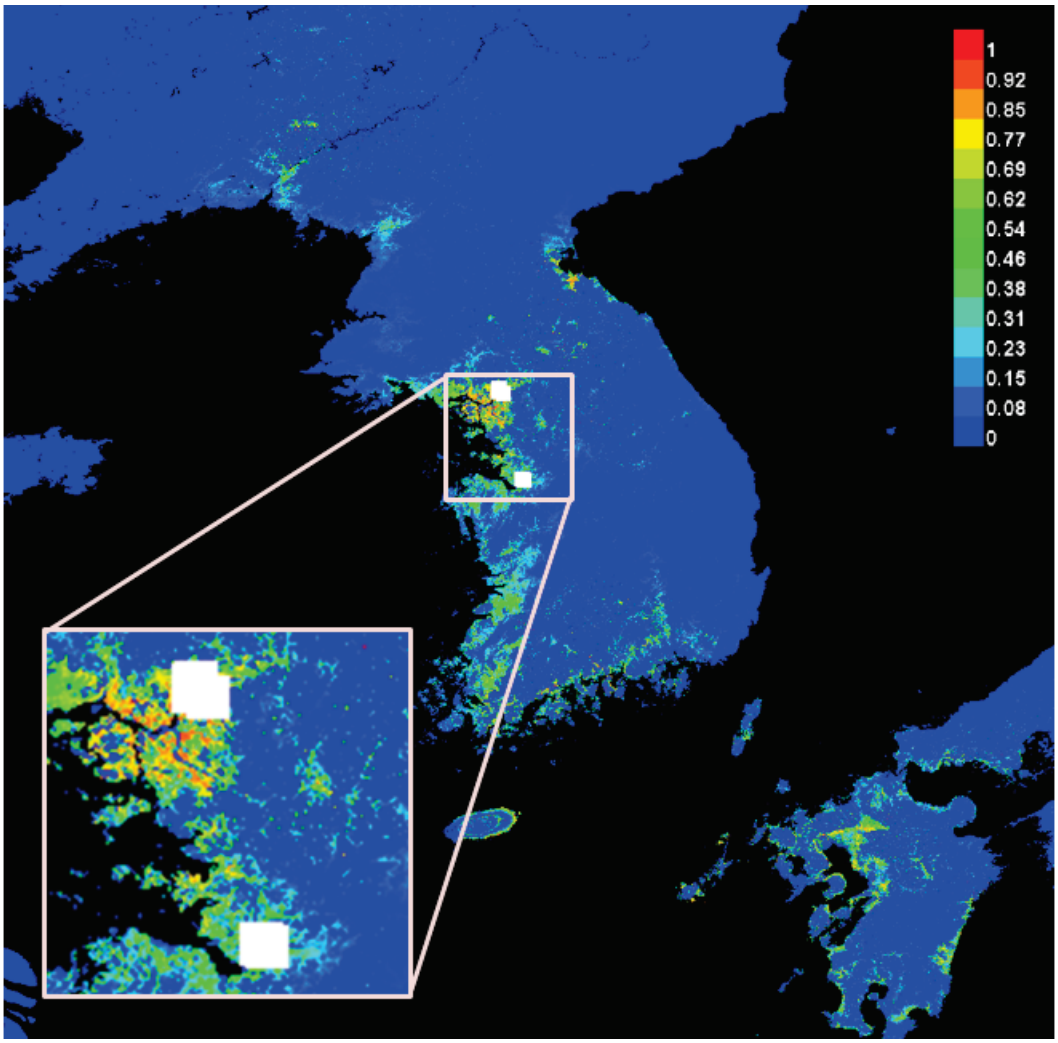

Figure 6. Maxent output, predicting likely presence of *Anopheles* breeding sites. The inset shows presence locations in large white squares.

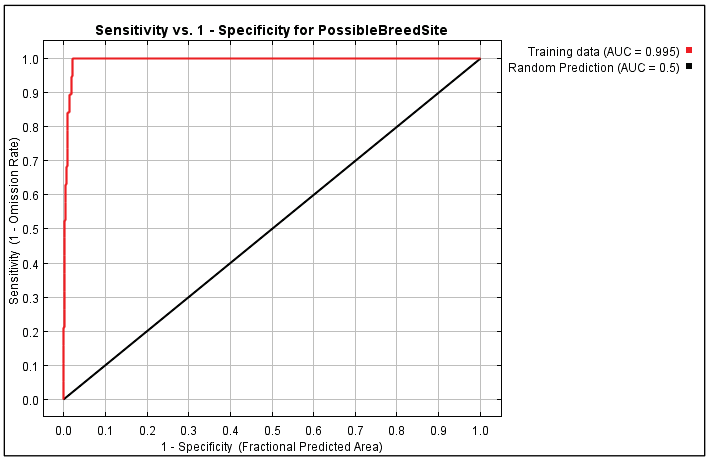

Figure 7. Specificity curve for Maxent output evaluation.

An additional output of the Maxent software is the percent contributions each variable made in the algorithm to build the models (Table 8). While all of the environmental variables contributed to the model to some extent, precipitation, elevation, and slope were the primary contributors. The contributions of these variables are similar to existing studies of the preferred environment of *Anopheles* mosquitos and reflect the choices the expert entomologists made in their field collections (Kim et al. 2011).

Table 8: Variable contributions for the Maxent model.

Variable	Percent contribution
Precipitation	53.1
Elevation	26.8
Slope	11.5
Population density	4.3
Mean NDVI	4.3

CONCLUSIONS: This TN describes the process used to predict the presence of breeding sites of mosquitos from the *Anopheles* genus in the Korean peninsula and surrounding areas. This method produces a map that gives planners insight into potential disease vector risks in the operational environment. A particular benefit of this method is that it can provide actionable information to planners where there is a lack of reliable surveillance information available. As military operations continue to evolve and integrate new types of information, tools like Maxent can be used to provide vital information to planners that supports military readiness through promoting health and effectiveness of the Warfighter.

REFERENCES

Center for International Earth Science Information Network. 2016. *Gridded Population of the World, Version 4 (GPWv4): Population Density.* NASA Socioeconomic Data and Applications Center (SEDAC). Accessed March 2019. doi:http://dx.doi.org/10.7927/H4NP22DQ.

Charlesworth, S. 2008. *Mosquitos: Purdue University Medical Entomology.* Accessed August 26, 2019. https://extension.entm.purdue.edu/publichealth/insects/mosquito.html.

Cho, S-Y., Y. Kong, S. M. Park, J. S. Lee, Y. A. Lim, S. L. Chae, W-G. Kho, J. S. Lee, J. C. Shim, and H-H. Shin. 1994. "Two vivax malaria cases detected in Korea." *Korean Journal of Parasitology* 32 (4): 281-284. doi:http://dx.doi.org/10.3347/kjp.1994.32.4.281.

Farr, T. G., P. A. Rosen, E. Caro, R. Crippen, R. Duren, S. Hensley, and M. Kobrick. 2007. "The shuttle radar topography mission." *Reviews of Geophysics* 45 (2). doi:https://doi.org/10.1029/2005RG000183.

Fukuda, M. M., M. Wojnarski, N. Martin, V. Zottig, and N. C. Waters. 2018. *Malaria in the Korean Peninsula: Risk Factors, Latent Infections, and the Possible Role of Tafenoquine, a New Antimalarial Weapon*. November 19. Accessed August 23, 2019. https://health.mil/News/Articles/2018/11/19/Malaria-in-the-Korean-Peninsula.

Funk, C., P. Peterson, M. Landsfeld, D. Pedreros, J. Verdin, S. Shukla, and G. Husak. 2015. "The Climate Hazards Infrared Precipitation with Stations - a new environmental record for monitoring extremes." *Scientific Data 2.* doi:150066. doi:10.1038/sdata.2015.66 .

Gaughan, A. E., F. R. Stevens, C. Linard, P. Jia, and A. J. Tatem. 2013. "High resolution population distribution maps for Southeast Asia in 2010 and 2015." *PLoS One* 8 (2).

Gorelick, N., M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore. 2017. "Google Earth Engine: Planetary-scale geospatial analysis for everyone." *Remote Sensing of Environment.*

Griffin, S. P. 2017. "Tactical-level Disease Vector Hazard Mapping using the Maximum Entropy (Maxent) Algorithm and ArcGIS Standard Operating Procedures (SOP)." SOP, Geospatial Research Laboratory, Engineer Research and Development Center. Accessed 2019.

Kim, H-C., L. M. Rueda, R. C. Wilkerson, D. H. Foley, W. J. Sames, S-T. Chong, P. V. Nunn, and T. A. Klein. 2011. "Distribution and larval habitats of Anopheles species in northern Gyeonggi Province, Republic of Korea." *Journal of Vector Ecology* (1): 124-134. doi:*https://doi.org/10.1111/j.1948-7134.2011.00149.x*.

Klein, T. A., G-C. Kim, W-J. Lee, L. M. Rueda, J. Sattobongkot, R. G. Moore, S-T. Chong, W. Sames, J. G. Pike, and R. C. Wilkerson. 2008. "Reemergence, persistence, and surveillance of vivax malaria and its vectors in the Republic of Korea." *Proceedings of the Sixth International Conference on Urban Pests.* 325-331.

Merow, C., M. J. Smith, and J. A. Silander, Jr. 2013. "A practical guide to MaxEnt for modeling species' distributions." *Ecography* 1058–1069. *https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1600-0587.2013.07872.x*.

Phillips, S. J., M. Dudík, and R. E. Schapire. 2004. "A maximum entropy approach to species distribution modeling." *In Proceedings of the Twenty-First International Conference on Machine Learning* 655-662. *http://rob.schapire.net/papers/maxent_icml.pdf*.

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. "Maximum entropy modeling of species geographic distributions." *Ecological Modelling* 190 (3-4): 231-259. Accessed July 29, 2019. *https://www.cs.princeton.edu/~schapire/papers/ecolmod.pdf*.

Radosavljevic, A., and R. P. Anderson. 2014. "Making better Maxent models of species distributions: complexity, overfitting, and evaluation." *Journal of Biogeography* 41: 629-643.

Rouse, J. W., R. W. Haas, J. A. Schell, and D. W. Deering. 1974. "Monitoring vegetation systems in the Great Plains with ERTS." *NASA Special Publication.* 351:309-317.

U.S. Department of the Army. 2015. "Army Health System Support Planning: Army Techniques Publication No. 4-02.55."

World Health Organization. 2019. *Malaria Fact Sheet.* 27 March. Accessed August 23, 2019. *https://www.who.int/news-room/fact-sheets/detail/malaria*.

NOTE: *The contents of this technical note are not to be used for advertising, publication, or promotional purposes. Citation of trade names does not constitute an official endorsement or approval of the use of such products.*