DOI: 10.24425/academiaPAN.2023.147606

ŚWIAT W KRZYWYM ZWIERCIADLE

GENOMART/SHUTTERSTOCK.COM

Drzewa rosnące na skraju lasu zwykle mają inny pokrój i wysokość niż te z jego wnętrza. Wybierając rośliny do badań, należy o tym pamiętać

Matematyka dysponuje narzędziami, które uznaje się za obiektywne, a obiektywność powinna być podstawowym kryterium naukowym. Czy metody statystyczne rzetelnie opisują świat?

Dominik Tomaszewski

Instytut Dendrologii PAN z siedzibą w Kórniku

Statystyka jest częścią metodologii badań, która wykorzystuje metody matematyczne do opisu i zrozumienia świata. Obraz całości, który uzyskujemy, powinien być wolny od spekulacji, kreacji, możliwie obiektywny – czyli po prostu prawdziwy, zgodny z rzeczywistością. Metody statystyczne mogą nam w tym znakomicie pomóc, jednak nadmiar zaufania do nich i ograniczona świadomość pułapek, w które można wpaść podczas analizy, powodują, że dość łatwo zamiast autentycznego świata pokazać przekłamaną i zniekształconą rzeczywistość.

Prowadząc badania, stawiamy hipotezy, które można potwierdzić lub zanegować. Do udzielenia odpowiedzi (prawda/fałsz) służą metody i kryteria oparte na analizie statystycznej, a do jej wykonania jest potrzebny odpowiedni zbiór danych. Gromadzenie danych to jednak dopiero początek drogi – po nim musi nastąpić proces ich weryfikacji, przetworzenia i właściwej analizy. Biolodzy najczęściej odwołują się do technik statystycznych z wykorzystaniem oprogramowania, które wykonuje obliczenia, informuje o wartościach parametrów korelacji, istotności różnic itp., jednak decyzja o doborze właściwej analizy oraz sprawdzenie założeń wstępnych (np. o wielkości próby, normalności rozkładu czy jednorodności wariancji) leży po stronie badacza. Zdarza się – niestety nader często – że czynności te są pomijane, co jest poważnym błędem i jako taki podważa cały ciąg rozumowania.

dr hab. Dominik Tomaszewski

Pracuje jako adiunkt w Instytucie Dendrologii PAN w Kórniku, zajmuje się mikromorfologią i systematyką roślin.

dominito@man.poznan.pl

Dokładność

Autentyczność obrazu zjawiska, który mamy nadzieję otrzymać, jest pochodną jakości danych. Jakości, czyli dokładności. Można mierzyć wysokość drzewa i przyrost jego grubości w ciągu roku, ale pomiar tych cech będzie wymagał innej dokładności. Określenie wysokości drzewa z precyzją do ułamka milimetra nie dość, że będzie niezwykle trudne technicznie, dodatkowo będzie wnosiło niską wartość merytoryczną. Podobnie jest z długością życia człowieka wyrażoną w sekundach czy z pomiarem współrzędnych geograficznych dużych obiektów z dokładnością do centymetrów. Taka dokładność nie jest potrzebna, by zlokalizować te duże obiekty w terenie.

Jednak pozornie wysoka dokładność nie jest wcale najgorszym z błędów metodycznych możliwych do popełnienia. Znacznie poważniejsza jest niereprezentatywność próby. Rzadko badamy całą grupę (czyli populację generalną), ponieważ statystyka wypracowała metody, które pozwalają na analizę takich zbiorów z wykorzystaniem tylko ich części (próby).

Liście nawet na tym samym pędzie mogą bardzo się różnić wielkością i kształtem – musimy to wiedzieć, jeśli chcemy przeprowadzić poprawną analizę tych cech

Jednak część, która trafia do analizy, musi dobrze reprezentować pierwotny, duży zbiór. Tylko taka próba pozwala odzwierciedlić populację generalną. Zwykle sprawdza się wówczas parametry rozkładu wartości cech, by stwierdzić, czy próba może zostać uznana za właściwą do dalszej analizy. W badaniach wysokości drzewostanu i budowy koron drzew w lesie próba złożona z osobników, które rosną na skraju (lub ich nieproporcjonalnie wysoki udział w naszej próbie), nie jest dobrym wyborem, wiadomo bowiem, że okrajek lasu charakteryzuje się występowaniem drzew o mniejszej wysokości i niżej rozgałęzionych niż te z wnętrza drzewostanu. Nadmiar takich drzew w badanej próbie zaburzyłby zatem obraz populacji generalnej, czyli naszego badanego lasu. Ten dość oczywisty przykład jasno pokazuje, że przy braku wystarczającej wiedzy można popełnić kardynalny błąd, wybierając do badań próbę niereprezentatywną, i nawet jeśli dalsze kroki analizy statystycznej zostaną wykonane prawidłowo, to i tak wnioski będą nieuprawnione.

Bardzo często dobór próby jest źródłem błędów w badaniach biologicznych, zwłaszcza na początku naukowego rozwoju badacza, gdy nie ma on jeszcze wielkiego doświadczenia. Wydawałoby się, że trywialny wybór terminu zbioru liści drzewa czy miejsce jego pobrania w koronie nie powinien mieć znaczenia. Nie jest jednak obojętne, czy wybierze się liście, które pojawiają się w maju, czy te wykształcone w czerwcu, z nasłonecznionej południowej części korony, czy z północnej, te, które rosną nisko i są dostępne z ziemi, czy te rosnące wysoko. Często jest potrzebna wiedza, by wybór był prawidłowy. Ważne jest, by wybrany podzbiór elementów autentycznie odzwierciedlał faktyczny rozkład cech w zbiorze ogólnym.

Zielniki

Jednym z warunków właściwego doboru próby jest losowość. Przy jego spełnieniu można domniemać, że faktycznie próba będzie reprezentatywna. Zazwyczaj jednak pełną losowość trudno osiągnąć. Za przykład mogą posłużyć pomiary okazów zielnikowych. Na ogół jest to sprasowana i wysuszona roślina, której towarzyszy etykieta z informacjami o jej przynależności systematycznej, o miejscu, czasie zbioru i zbieraczu. Okazy zielnikowe gromadzi się i przechowuje w wyspecjalizowanych jednostkach naukowych zwanych zielnikami lub herbariami. Okazy zielnikowe są cennym źródłem danych o roślinach, ponieważ proces suszenia dość dobrze zachowuje ich budowę, w dodatku można z nich pozyskiwać materiał genetyczny do badań różnego rodzaju. Herbaria obecnie są miejscem przechowywania setek milionów okazów roślin z całego świata z wszystkich grup systematycznych i choć ich główna część została zebrana w czasie ostatniego stulecia, to jednak zakres czasowy jest znacznie szerszy – najstarsze zachowane okazy zielnikowe liczą sobie blisko 500 lat. Jest to więc przebogate źródło danych biologicznych.

Okazy zielnikowe są często wykorzystywane do badań botanicznych, np. w analizach biometrycznych, które bazują na pomiarach cech (takich jak wymiary liści, owoców czy elementów kwiatów). Na pierwszy rzut oka nie powinno to budzić zastrzeżeń. Jednak i tu można napotkać trudności metodyczne. Po pierwsze – wspomniana już losowość. Idealna próba badawcza powinna składać się z osobników zebranych stochastycznie, tymczasem zbieracz w terenie zwykle narusza to kryterium, ponieważ wybiera rośliny, które z jakiegoś powodu wzbudziły jego zainteresowanie i dlatego zdecydował się na wybór właśnie tego, a nie innego osobnika spośród dziesiątków tam obecnych (np. wyjątkowo mały albo nader okazały, wygodny do zbioru i suszenia, o rzadkim zabarwieniu kwiatów czy liści, nietypowym pokroju itd.). Tym samym

ŹRÓDŁO: GBIF

Dane o występowaniu pokrzywy zwyczajnej we wschodnich Niemczech i zachodniej Polsce na tej mapie ujawniają nie tyle autentyczny obraz rozmieszczenia tego gatunku, ile raczej istnienie bardzo różnych poziomów dostępności danych o bioróżnorodności

zostaje wprowadzony element nielosowości wyboru. To z kolei oznacza, że rozkład wartości cech w badanej próbie zapewne nie będzie odzwierciedlać rozkładu wartości w populacji generalnej.

W przypadku okazów zielnikowych trzeba zdawać sobie sprawę z jeszcze innego niebezpieczeństwa. Jak wiadomo, tkanki rośliny zawierają dużo wody, a ta w procesie suszenia zostaje usunięta. Ma to dalsze konsekwencje w aspekcie oceny cech. Zwykle zmienia się zabarwienie liści czy kwiatów, czasem tak bardzo, że doświadczeni zbieracze, zdając sobie z tego sprawę, odnotowują pierwotną barwę na etykiecie zielnikowej. Ponadto, skoro tkanki zostały odwodnione, zmieniają się także rozmiary organów. I tak się faktycznie dzieje. Przy badaniu liści około 20 gatunków z różnych grup stwierdziliśmy, że liście traciły 52–86 proc. masy, a jednocześnie ich powierzchnia zmniejszyła się o 3,5–15,2 proc. (z literatury wiadomo, że spadek może być jeszcze większy). Co więcej, wykorzystując metody kwantyfikacji kształtu z użyciem eliptycznych współczynników Fouriera, można stwierdzić, że nie zostaje zachowana pierwotna forma liści. Blaszka liściowa wprawdzie po wysuszeniu nie wygląda zupełnie inaczej i nie zmienia się nie do poznania: botanik nadal poprawnie zidentyfikuje dany gatunek. Ale ktoś, kto nie weźmie pod uwagę takich zmian, a będzie analizował zbiór danych, do którego trafiły pomiary liści świeżych i zasuszonych, popełni błąd metodyczny, który rzecz jasna będzie rzutował na wyniki i ich późniejszą interpretację.

Wielkie zbiory danych

W wielu sytuacjach rozwiązaniem problemów związanych z wielkością i losowością próby byłaby analiza *big data*, czyli wielkich zbiorów danych. Obecne narzędzia informatyczne i dostęp do dużych ilości danych z różnych źródeł pozwalają na pracę ze zbiorami nie dziesiątek czy setek danych, lecz setek tysięcy i milionów. W biologii świetnym tego przykładem jest dostęp do danych o bioróżnorodności. Dzięki uruchomionym stosunkowo niedawno procesom digitalizacji kolekcji biologicznych z każdym miesiącem przybywa łatwo dostępnych danych o występowaniu organizmów. Mimo to taki ogromny zbiór i tak nie jest doskonały. Mając na mapie nawet miliony punktów wskazujących na miejsce występowania, nic nie wiemy o tych milionach, które się na niej nie znalazły. Jako przykład mogą posłużyć obserwacje występowania pokrzywy zwyczajnej (*Urtica dioica*), gatunku bardzo pospolitego w Polsce i Niemczech. W rozmieszczeniu pokrzywy zaskakuje ogromna liczba miejsc występowania na zachód od Odry, a mała na wschód od niej. Każda z obserwacji jest prawdziwa, dlaczego więc całościowy obraz nie jest? Kluczem jest niedoskonałość danych (*bias*). Ponieważ baza danych o bioróżnorodności, w tym przypadku największa: Global Biodiversity Information Facility, nie dysponuje odpowiednią liczbą obserwacji z Polski, rejon ten jest niedoszacowany. Rzecz w tym, że procesy digitalizacji i przekazywania danych do ogólnoświatowych baz nie przebiegają równie szybko i wydajnie w różnych krajach i dlatego bazy są „zbiasowane", czyli z jakiegoś powodu struktura danych jest zaburzona. Z czasem to się powinno zmienić i wówczas *big data* pozwolą na jeszcze lepszy, bardziej zbliżony do autentycznego opis świata.

Trudno ocenić, z iloma tego typu pułapkami ma do czynienia biolog w swojej pracy. Kolejne, na które trzeba uważać, to niedopasowanie analizy do typu danych i do pytania badawczego czy błędna interpretacja poprawnych wyników. Zbierając dane do dalszej analizy, trzeba pamiętać o pułapkach, których można uniknąć. ■

Chcesz wiedzieć więcej?

Bąk J., *Statystycznie rzecz biorąc*, 2020.

Huff D., *Dlaczego statystyki kłamią*, 2023.

Stephens-Davidowitz S., *Wszyscy kłamią: big data, nowe dane i wszystko, co Internet może nam powiedzieć o tym, kim naprawdę jesteśmy*, 2019.