

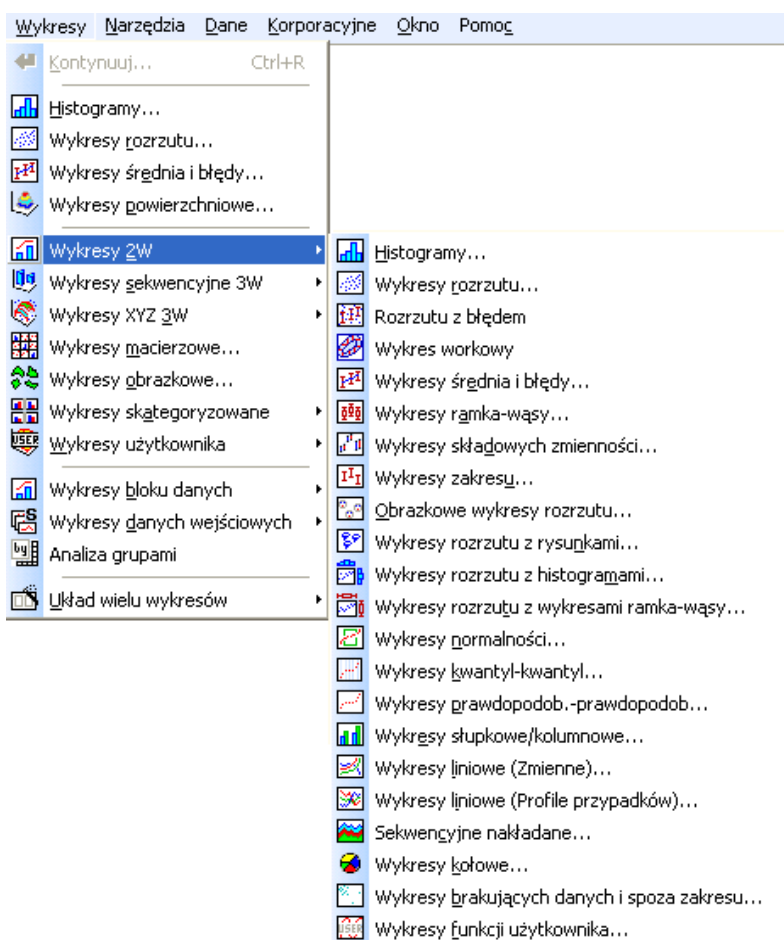


WIZUALIZACJA DANYCH JAKO UZUPEŁNIENIE METOD ANALITYCZNYCH

Krzysztof Suwada, StatSoft Polska Sp. z o.o.

Techniki graficzne stanowią efektywny sposób prezentacji i przekazywania informacji. Dobrze zaprojektowany wykres jest w stanie zastąpić setki, a nawet tysiące liczb. Ponadto różne techniki graficzne mogą stanowić potężne analityczne narzędzia do eksploracji danych i sprawdzania hipotez.

STATISTICA zawiera obszerny wybór metod graficznych, służących zarówno do analizy danych, jak i prezentacji wyników.



Rys. 1. Lista wykresów.



Wszystkie wykresy dostępne w programie *STATISTICA* zawierają szereg wbudowanych interaktywnych technik analitycznych oraz szeroki zakres narzędzi dostosowywania, umożliwiających użytkownikowi interaktywne sterowanie prawie wszystkimi aspektami wykresu.

Podczas prezentacji przedstawione zostaną przykładowe metody, którymi można posługiwać się podczas analizy danych, otrzymując ciekawe wykresy i zestawienia.

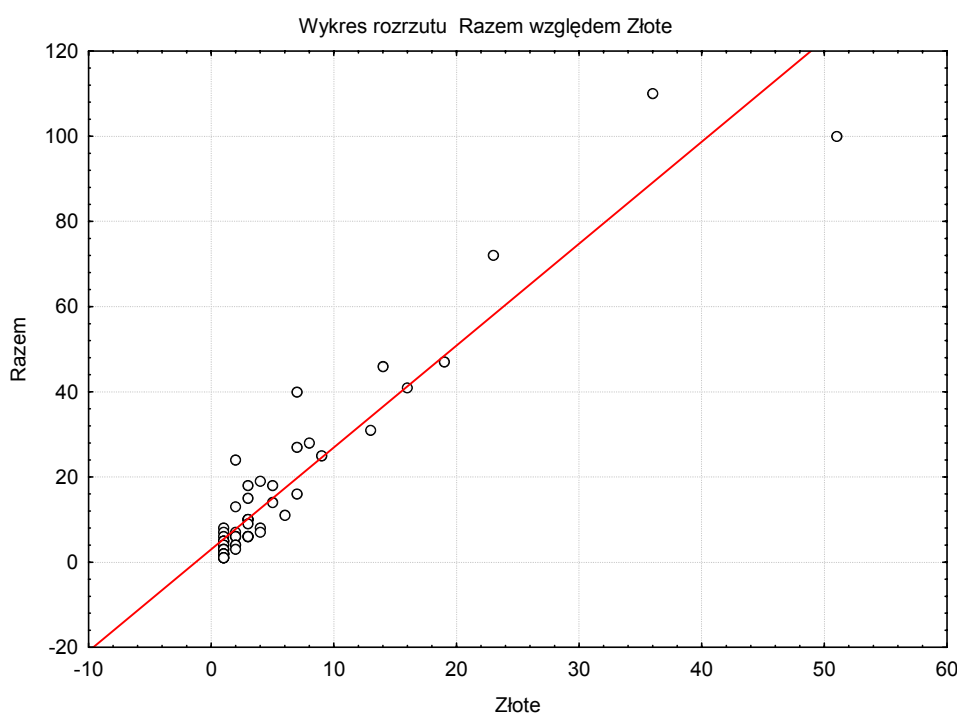
Wykresy wykonane w programie *STATISTICA* można bardzo łatwo eksportować do popularnych formatów graficznych, takich jak: wmf, emf, JPG, tiff czy png. Wykresy utworzone w programie można także osadzać w innych aplikacjach, dzięki czemu możliwa jest ich późniejsza edycja, można je wtedy wygodnie skalować bez utraty jakości.

Identyfikacja obiektów i segmentacja

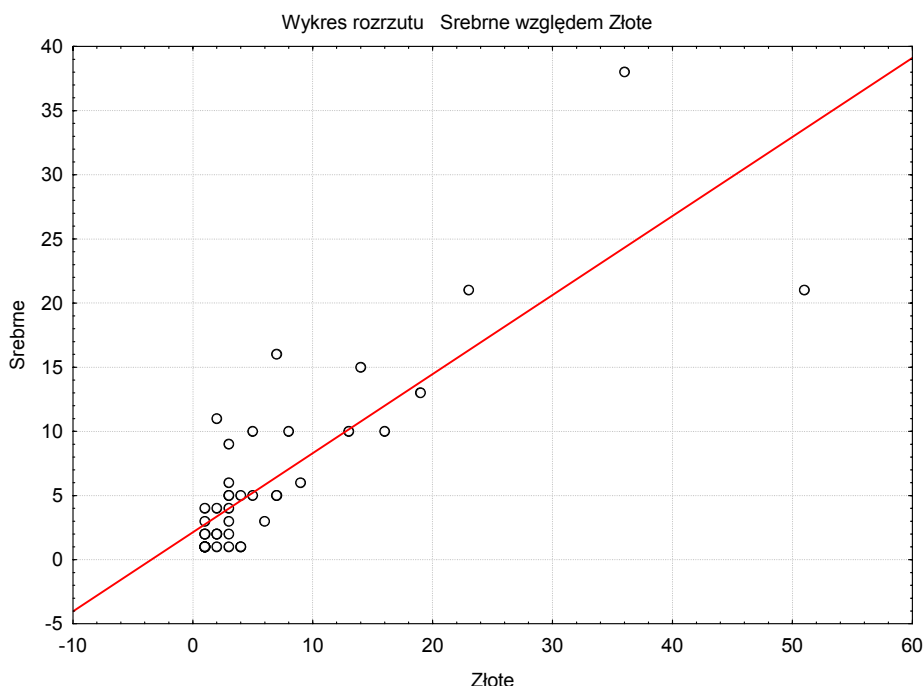
Jedną z podstawowych technik analizy i zwykle jedną z pierwszych jest tworzenie prostych charakterystyk danych, jak: średnia, mediana czy odchylenie standardowe. Taki opis zestawem liczb jest cenny, ale nie daje analitykowi pełnej informacji o zbiorze danych. Bardzo cennych informacji dostarczają różnego rodzaju wykresy rozrzutu.

W naszym przykładzie wykorzystamy dane o klasyfikacji medalowej poszczególnych państw prowadzonej przez MKOI. Dane zawierają informacje o ogólnej licznie zdobytych medali we wszystkich konkurencjach – złotych, srebrnych i brązowych. Na tej podstawie wyznaczany jest ranking państw.

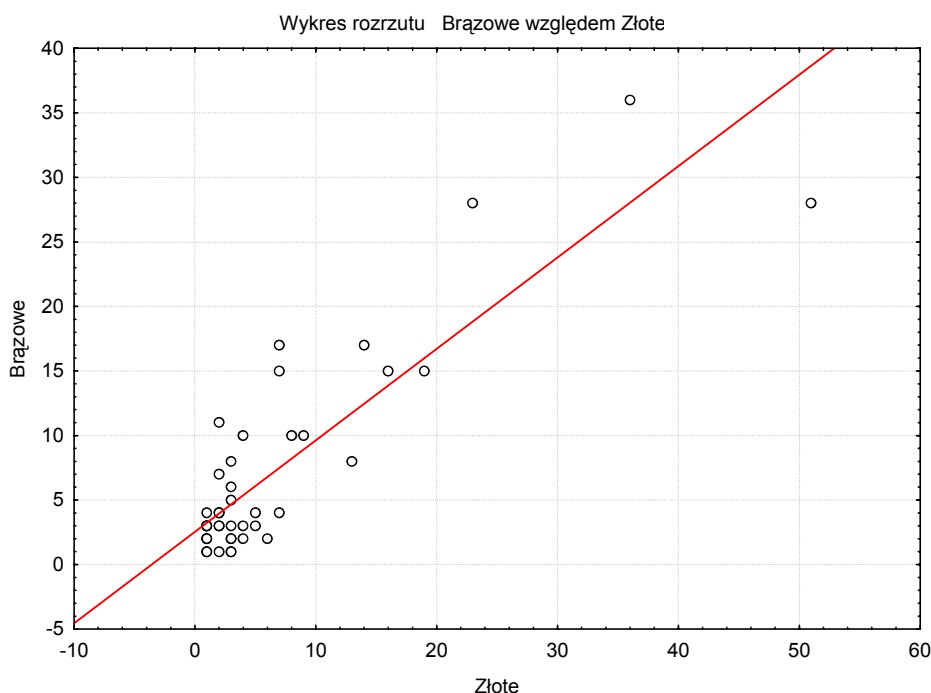
Aby dokonać wstępnej eksploracji zbioru danych, utwórzmy wykresy rozrzutu.



Rys. 2. Sumaryczna liczba medali a złote medale.



Rys. 3. Liczba srebrnych a liczba złotych medali.




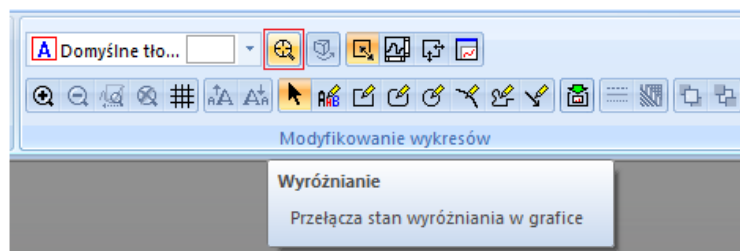
Rys. 4. Liczba brązowych a liczba złotych medali.

Utworzone wykresy dostarczają informacji na temat zależności między poszczególnymi parami zmiennych. Na każdym z wykresów łatwo zauważyć kilka obserwacji nietypowych, które znajdują się w okolicy prawego górnego rogu wykresu.

Program *STATISTICA* dostarcza bardzo użytecznego narzędzia pozwalającego bardzo łatwo dokonać identyfikacji poszczególnych przypadków na wykresie. Aby sprawdzić,

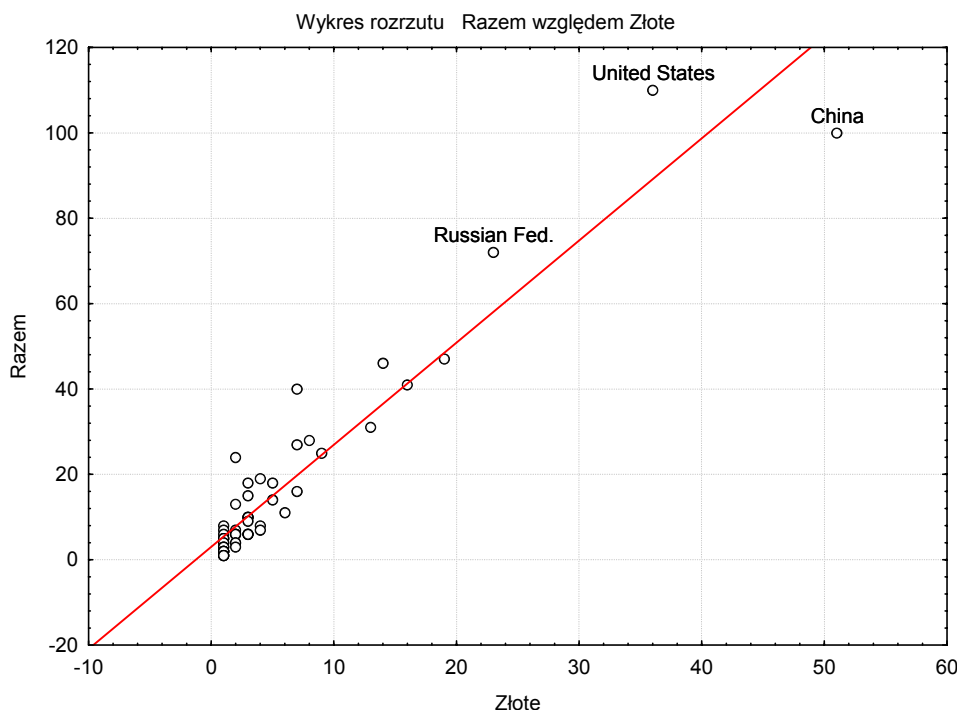


które państwa odstają od pozostałych, wykorzystamy narzędzie *Wyróżnianie*. W tym celu, mając aktywny wykres, klikamy ikonkę , którą na poniższym rysunku zaznaczono czerwonym kwadratem.



Rys. 5. Wyróżnianie.

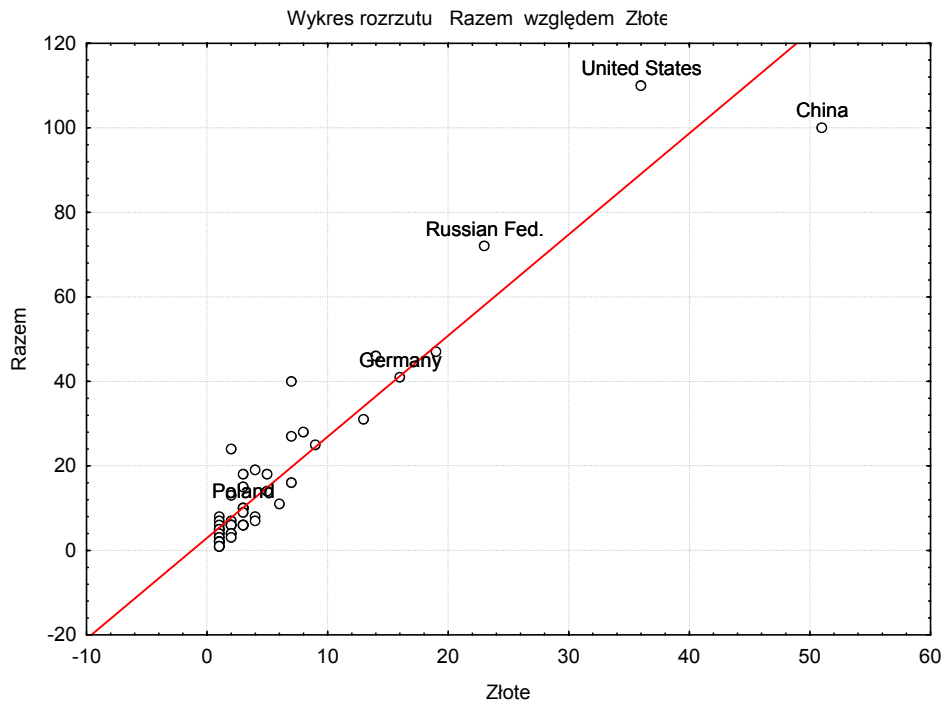
W oknie, które się pojawi, wybieramy *Etykietuj* oraz celownik *Ramka*. Otaczamy ramką interesujące nas punkty i klikamy przycisk *Zastosuj*.



Rys. 6. Efekt wyróżniania.

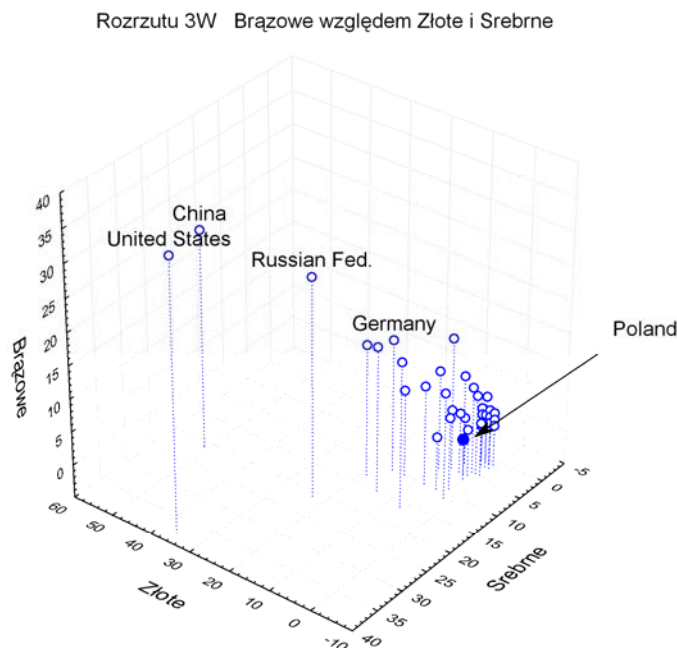
W ten prosty sposób udało nam się błyskawicznie zidentyfikować obiekty, które wyraźnie odstają od pozostałych. Interesującą z naszego punktu widzenia jest także informacja, gdzie wśród tych wszystkich punktów znajduje się Polska i np. jeden z naszych sąsiadów – Niemcy.

Aby zobaczyć, gdzie na wykresie znajdują się odpowiadające im punkty, przechodzimy do arkusza i oznaczamy interesujące nas przypadki jako *Etykietowane*. Jak widać, przypadki zaznaczone na wykresie mają już ustawioną taką właściwość.



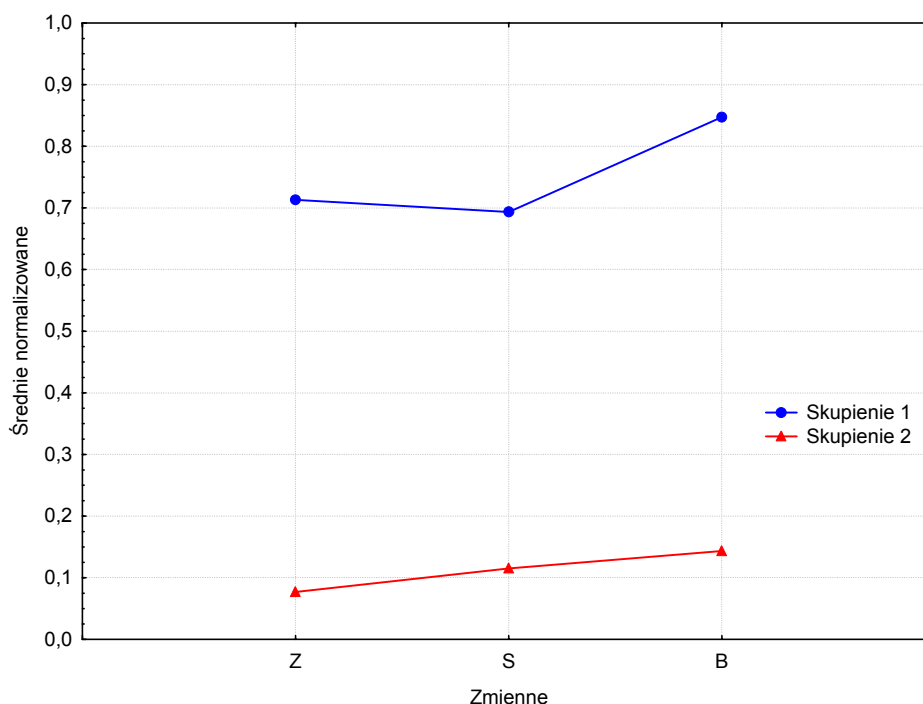
Rys. 7. Polska i kraje prowadzące w klasyfikacji medalowej.

Ze względu na małą liczbę zmiennych, możemy także zobaczyć, jak poszczególne punkty układają się na wykresie 3D, wybierając jako poszczególne osie liczby zdobytych medali. Zanim wykonamy wykres, oznaczymy przypadek odpowiadający Polsce nie jako *Etykietywany*, ale *Zaznaczony*. Na trójwymiarowym wykresie rozrzutu będzie on zaznaczony wypełnioną na niebiesko kropką.



Rys. 8. Polska i kraje wiodące w klasyfikacji medalowej w 3D.

Wykonane rysunki sugerują, że mamy do czynienia z dwoma znacząco różniącymi się grupami państw. Nasze przypuszczenia możemy sprawdzić, wykonując segmentację, np. metodą k-średnich lub metodą EM (obie są dostępne w programie *STATISTICA*). Wyniki segmentacji potwierdzają nasze przypuszczenia. Analiza ujawniła istnienie dwóch istotnie różnych statystycznie segmentów, do pierwszego skupienia należą: Chiny, USA oraz Rosja. W drugim znajdują się wszystkie pozostałe kraje. Wyniki analizy prezentuje wykres.



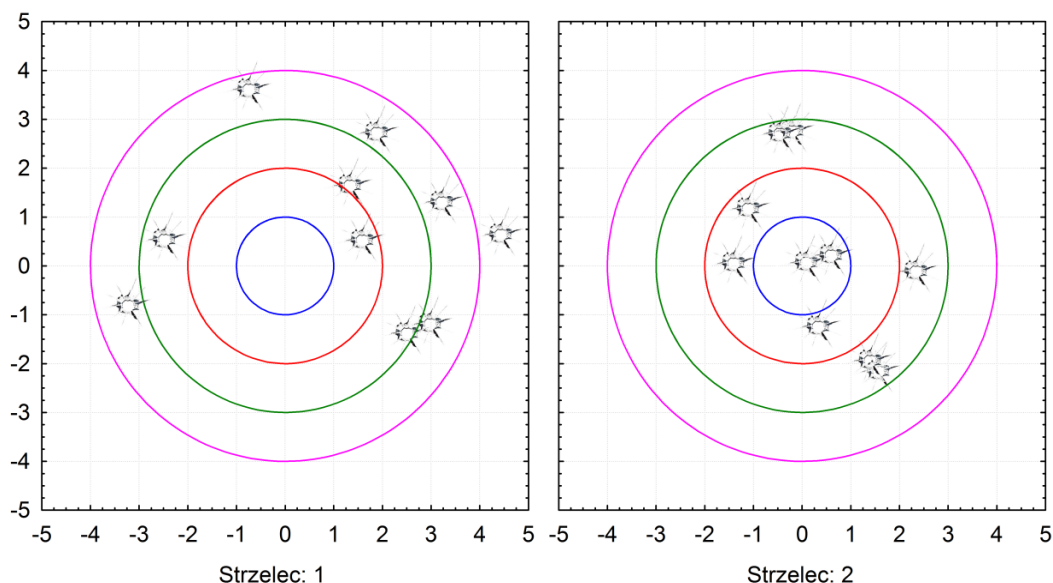
Rys. 9. Średnie skupień po segmentacji.

Wizualizacja wyników

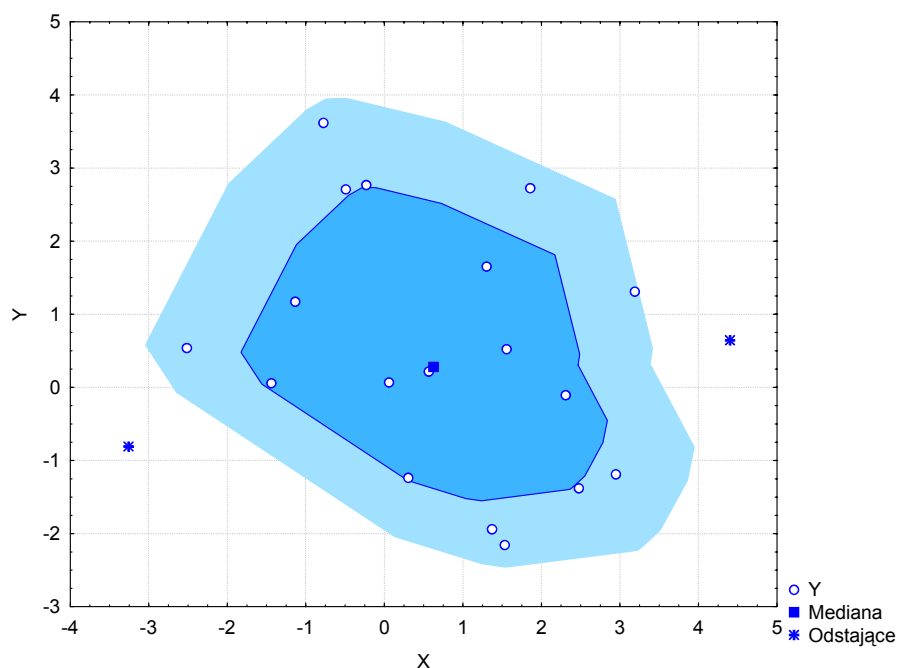
Obecnie praktycznie w każdej dziedzinie życia wykorzystuje się mniej lub bardziej zaawansowane metody analizy danych i ich wizualizacji. Jak widzimy, można je wykorzystywać także w sporcie. W kolejnym przykładzie zademonstrujemy możliwości wizualizacji danych w programie *STATISTICA* na przykładzie wyników dwóch strzelców.

Założmy, że mamy dane dotyczące współrzędnych trafień w tarczę dla dwóch strzelców, jako środek tarczy przyjmujemy punkt (0,0). Na początek zobaczmy, jak wyglądały tarcze obu strzelców tuż po strzelaniu. W programie *STATISTICA* bardzo łatwo ten cel osiągnąć, korzystając ze skategoryzowanego wykresu rozrzutu (rys. 10).

Widzimy teraz dokładnie, gdzie trafił dany strzelec. Jako dodatkowy efekt, zamiast standardowego znacznika punktu został użyty obrazek dziury po kuli. Kolejnym krokiem może być bardziej wnikliwa analiza wyników strzelania, wykonana np. z wykorzystaniem wykresu workowego, tym razem już ze standardowymi znacznikami (rys. 11).



Rys. 10. Wyniki strzelców.



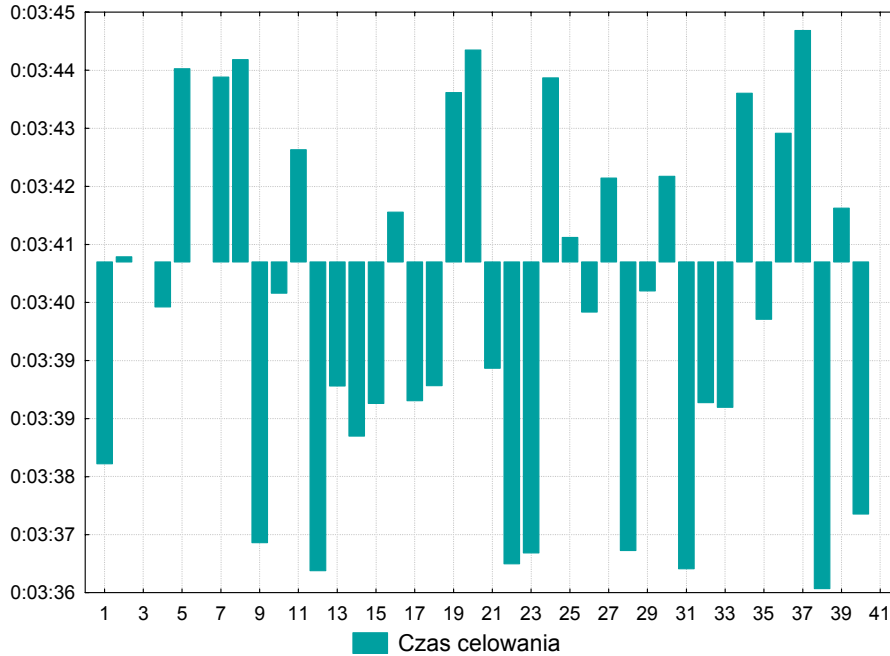
Rys. 11. Wykres workowy.

Wykres workowy w przystępny sposób pokazuje nam, gdzie najczęściej powinien trafiać strzelec, z wykorzystaniem dwuwymiarowego uogólnienia Tukeya jednowymiarowego wykresu ramka-wąsy dla identyfikacji rozkładu (i wartości odstających) w przestrzeni dwuwymiarowej. Jak widać, jest sporo obserwacji odstających, co może sugerować rozregulowanie przyrządów celowniczych lub inny problem ze sprzętem.

W dzisiejszych czasach w celu osiągnięcia jak najlepszych wyników niemal wszystkie aspekty rywalizacji podlegają optymalizacji. Jednym z czynników, który może wpłynąć na celność, jest czas celowania. Zobaczmy, jak to wygląda w naszym przypadku. Na podstawie

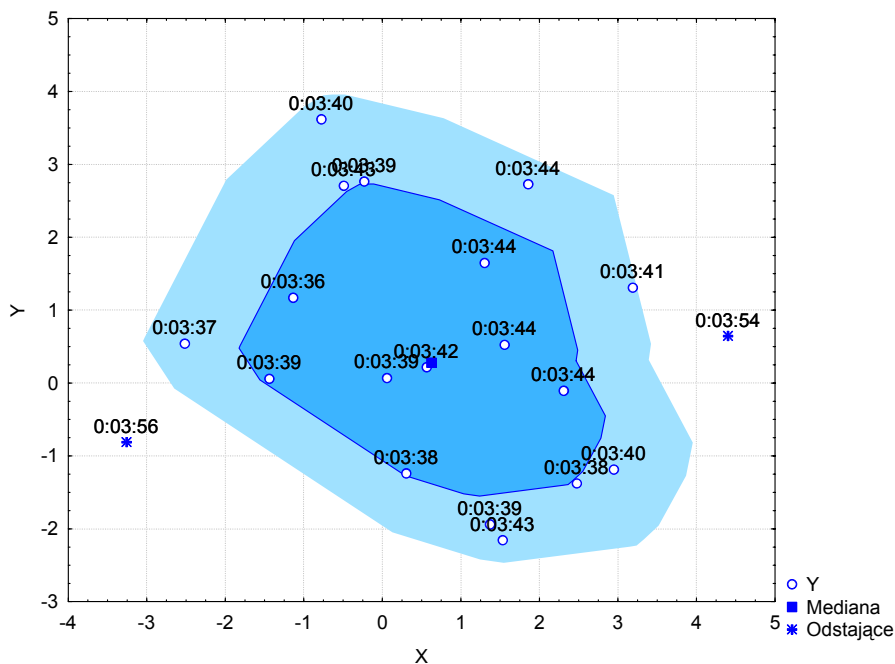


40 strzałów otrzymano oszacowanie średniej czasu celowania na poziomie około 00:03:41. Zobaczmy, jak na wykresie słupkowym wyglądają pozostałe czasy celowania. Aby ułatwić sobie zadanie interpretacji, zmieniamy także punkt odniesienia, przesuując „zero”.



Rys. 12. Analiza czasu celowania.

Jak widać, w niektórych przypadkach czas celowania jest o kilkanaście setnych sekundy krótszy lub dłuższy niż średni. Zobaczmy, czy ma to wpływ na celność, wykonując ponownie wykres workowy i etykietując punkty wartościami czasu celowania.



Rys. 13. Czas celowania a obserwacje odstające.

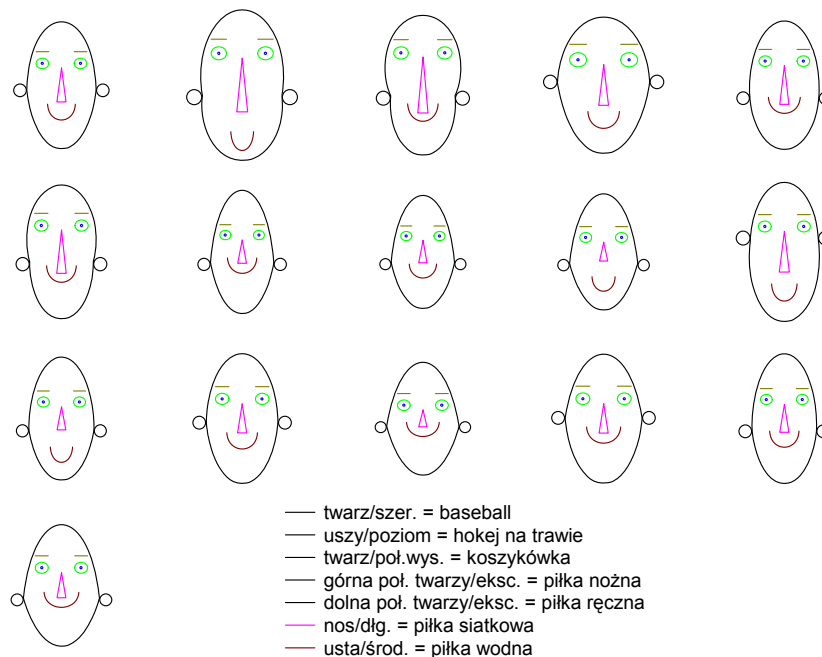


Jak widać, dwie odstające obserwacje mają czasy celowania o kilkanaście setnych sekundy wyższe od średniej. Można podejrzewać, że dłuższy czas celowania – nawet o kilkanaście setnych sekundy niekorzystnie wpływa na celność. Oczywiście ilość danych zebranych na tym etapie jest niewystarczająca, ale wskazuje jeden z możliwych kierunków dalszych badań.

Wykonaliśmy tylko kilka prostych wykresów, ale łatwo zauważyć, że ilość przekazywanych przez nie informacji jest dosyć duża, a sposób ich prezentacji znacznie ułatwia ich analizę.

Zależności geograficzne

Często zdarza się, że dane, które chcemy analizować, dotyczą pewnego obszaru geograficznego, np. kraju, województwa czy powiatu. Zwykle tego typu dane składają się z większej liczby zmiennych i nie da się ich przedstawić na wykresie rozrzutu. Możliwym rozwiązaniem jest przedstawienie ich z wykorzystaniem twarzy Chernoffa. W programie *STATISTICA* dostępne są one w menu *Wykresy obrazkowe*. Wynik wizualizacji przedstawiony został poniżej na rys. 14.

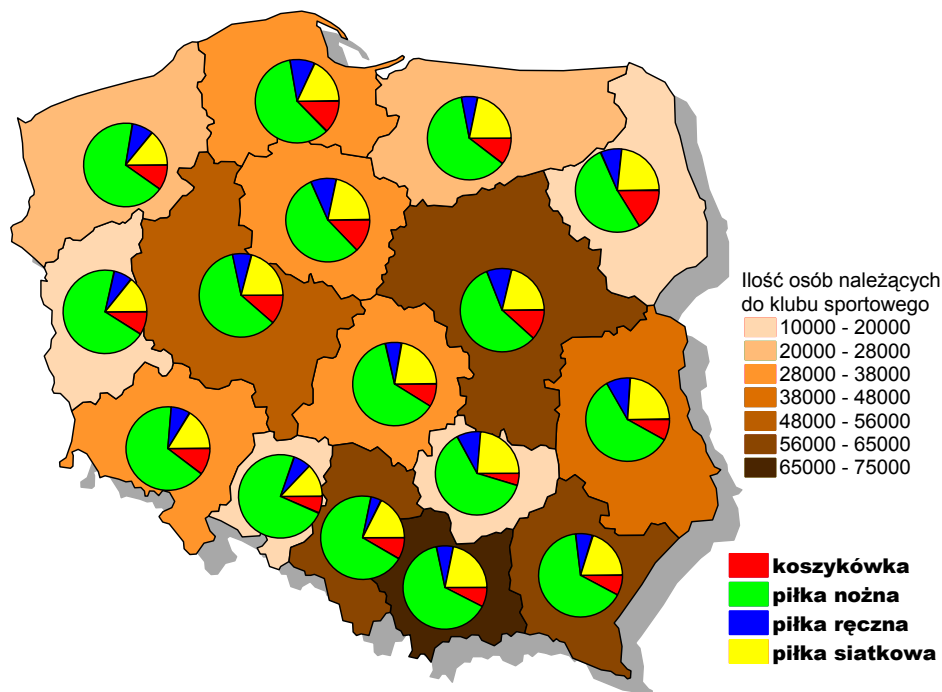


Rys. 14. Twarze Chernoffa.

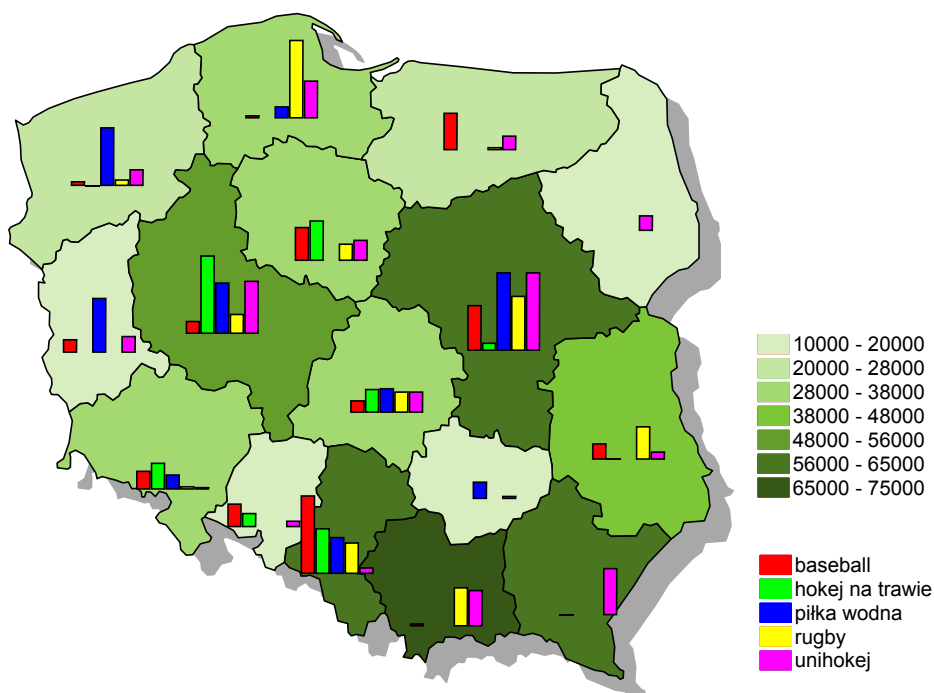
Twarze Chernoffa to jeden z możliwych sposobów wizualizacji danych wielowymiarowych, jednak aby analiza takich danych była kompletna, nie można zapomnieć o wizualizacji danych na odpowiednich mapach. Przedstawienie danych tylko w postaci zwykłej tabelki może doprowadzić do pominięcia pewnych istotnych zależności wynikających właśnie z położenia geograficznego.



W środowisku *STATISTICA* do wizualizacji służy dodatek Mapy, który można pobrać ze strony www.statsoft.pl. W kolejnym przykładzie posłużymy się danymi o liczbie osób uprawiających poszczególne dyscypliny sportowe w klubach na terenie Polski.



Rys. 15. Mapy i wykresy kołowe.



Rys. 16. Mapy i wykresy słupkowe.



Z rys. 15 jesteśmy w stanie odczytać informację o ogólnej liczbie osób należących do klubów sportowych (kolor tła mapy), zobaczyć, jak ta liczba rozkłada się na terytorium całego kraju. Dodatkowo w postaci wykresów kołowych przedstawiono strukturę popularności wybranych dyscyplin. Dzięki wykresom można bez trudu porównać ich popularność w poszczególnych województwach.

Inną ciekawą funkcjonalnością jest możliwość nakładania na mapę wykresów słupkowych, w których słupki skalowane są względem danej kolumny (rys. 16). Umożliwia to graficzną ocenę, w którym województwie dana dyscyplina sportowa jest najbardziej popularna.

Ilość informacji przedstawionych na tej jednej mapie i ich czytelność jest czymś nieosiągalnym, gdy korzystamy z tabeli, czy nawet zestawu tabel. Warto więc pamiętać także o tym sposobie wizualizacji danych.

Literatura

1. G. Harańczyk, *Metody wizualizacji danych*, Materiały kursowe StatSoft Polska, 2009.
2. <http://Gus.pl>.
3. <http://results.beijing2008.cn/WRM/ENG/INF/GL/95A/GL0000000.shtml>.