INSTYTUT ZOOTECHNIKI
PAŃSTWOWY INSTYTUT BADAWCZY

Rozprawa doktorska

Dziedzina: Nauki rolnicze

Dyscyplina naukowa: Zootechnika i Rybactwo

Ocena wartości hodowlanej bydła przy zastosowaniu modelu jednostopniowego

Mgr inż. Dawid Słomian

Praca doktorska wykonana pod kierunkiem
promotor prof. dr hab. Joanna Szyda
oraz
promotor pomocniczy dr inż. Kacper Żukowski

Kraków, 2025 rok

Serdeczne podziękowania składam Pani prof. dr hab. Joannie Szydzie za merytoryczne wsparcie, cenne wskazówki oraz życzliwość okazywaną na każdym etapie realizacji badań. Dziękuję za inspirujące rozmowy, wnikliwe uwagi oraz konsekwentne motywowanie do doskonalenia warsztatu naukowego. Pani doświadczenie, cierpliwość i zaangażowanie miały kluczowe znaczenie dla powstania tej pracy.

Serdeczne podziękowania kieruję do Pana dr Kacpra Żukowskiego za pomoc w prowadzeniu badań, trafne sugestie oraz wsparcie merytoryczne w rozwiązaniu problemów badawczych. Dziękuję za poświęcony czas, dostępność i konstruktywne uwagi, które znacząco przyczyniły się do podniesienia jakości niniejszej pracy.

Wyrazy wdzięczności składam Pracownikom Zakładu Hodowli Bydła w Instytucie Zootechniki za umożliwienie realizacji badań oraz życzliwe wsparcie organizacyjne i techniczne.

Z całego serca dziękuję Rodzicom i Rodzinie za nieustanne wsparcie, wyrozumiałość i motywację w trakcie całej drogi naukowej, Dziękuję za wiarę we mnie, cierpliwość oraz pomoc w chwilach zwątpienia. Szczególne podziękowania kieruje do mojej Narzeczonej Kamili za codzienną obecność, zrozumienie i ogromną cierpliwość, które dodawały mi sił i pomagały konsekwentnie dążyć do celu.

Oświadczenie promotora rozprawy doktorskiej

Oświadczam, że niniejsza rozprawa doktorska została przygotowana pod moim kierunkiem i stwierdzam, że spełnia ona warunki do przedstawienia jej w postępowaniu o nadanie stopnia naukowego.

Data ..... 15.12.2025 .....

Podpis promotora ..... M. Snyda .....

Oświadczenie autora rozprawy doktorskiej

Świadom odpowiedzialności prawnej oświadczam, że:

- niniejsza rozprawa doktorska została przygotowań przeze mnie samodzielnie pod kierunkiem Promotora i Promotora pomocniczego i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

- przedstawiona rozprawa doktorska nie była wcześniej przedmiotem procedur związanych z uzyskaniem stopnia naukowego.

- niniejsza wersja rozprawy doktorskiej jest tożsama z załączoną na płycie CD wersją elektroniczną.

Data ..... 15. 12. 2025 .....

Podpis autora ..... D. Stomczyk .....

Spis treści

Wykaz prac naukowych wchodzących w skład rozprawy

P1. **Słomian, D.**, Żukowski, K., Szyda, J. (2023). Heterogeneity in convergence behaviour of the single-step SNP-BLUP model across different effects and animal groups. Genetics Selection Evolution, 55(1). https://doi.org/10.1186/s12711-023-00856-5

MEIN: 100 IF: 4.2

Wkład autorski Słomian D.: udział w opracowywaniu metodyki, przygotowanie danych, wykonanie analiz, wiodący udział w przygotowaniu manuskryptu oraz współredagowaniu powstałej publikacji.

P2. **Słomian, D.**, Żukowski, K., Szyda, J. (2025). A comparison of genomically enhanced breeding values predicted by different single-step approaches. Annals of Animal Science. https://doi.org/10.2478/aoas-2025-0088

MEIN: 140 IF: 2.2

Wkład autorski Słomian D.: udział w opracowywaniu metodyki, przygotowanie danych, wykonanie analiz, wiodący udział w przygotowaniu manuskryptu oraz współredagowaniu powstałej publikacji.

P3. **Słomian, D.,** Jakimowicz, M., Suchocki, T., Szyda, J. (2025). Comparison of BLUPF90IOD3 and MiXBLUP implementations of the single-step model applied to the Polish national dairy cattle evaluation. PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-8398690/v1]

MEIN: 0 IF: 0

Wkład autorski Słomian D.: udział w opracowywaniu metodyki, udział w przygotowaniu danych, udział w wykonaniu analiz, wiodący udział w przygotowaniu manuskryptu oraz współredagowaniu powstałej publikacji.

P4. **Słomian, D.,** Vandenplas, J., Ten Napel, J., Żukowski, K., Skarwecka, M., Szyda, J. (2025). Modeling missing parents in single-step test-day SNP-BLUP evaluation of dairy cattle. [Preprint]. bioRxiv. https://doi.org/10.64898/2025.12.02.691779

MEIN: 0 IF: 0

Wkład autorski Słomian D.: udział w opracowywaniu metodyki, przygotowanie danych, wykonanie analiz, wiodący udział w przygotowaniu manuskryptu oraz współredagowaniu powstałej publikacji.

Sumaryczna punktacja MEIN: 240 punktów

Sumaryczny Impact Factor: 6.4

Źródło finansowania badań

Podstawą niniejszej dysertacji są wyniki badań w ramach dotacji statutowych Instytutu Zootechniki – Państwowego Instytutu Badawczego, pt. „Wykorzystanie alternatywnej populacji referencyjnej w genomowej ocenie wartości hodowlanej bydła rasy holsztyńsko-fryzyjskiej„ o numerze rejestracyjnym 04-10-10-11 oraz pt. „Analiza wpływu grup genetycznych oraz metafounders na wyniki modelu jednostopniowego dla cech produkcyjnych w ocenie wartości genetycznej bydła” o numerze rejestracyjnym 04-10-14-11.

Wykaz stosowanych skrótów

1. **APY** (Algorithm of Proven and Young) - model G-BLUP z użyciem zestawu osobników rdzeniowych, w celu uniknięcia odwracania całej macierzy spokrewnienia genomowego
2. **BLUP** (Best Linear Unbiased Prediction) - najlepsza liniowa nieobciążona predykcja
3. **DRP** (Deregressed Proof) - wartości hodowlane poddane deregresji
4. **EDC** (Effective Daughter Contribution) - efektywna liczba córek
5. **G-BLUP** (Genomic BLUP) - genomowa najlepsza liniowa nieobciążona predykcja
6. **GEBV** (Genomic Estimated Breeding Value) - genomowa oszacowana wartość hodowlana
7. **GG** (Genetic Groups) - grupy genetyczne
8. **GT** (Podejście GT) - algorytm szacowania efektów modelu G-BLUP z użyciem macierzy $\boldsymbol{T}$
9. **MACE** (Multiple-trait Across Country Evaluation) - metoda wielocechowej oceny międzynarodowej wartości hodowlanej
10. **MAF** (Minor Allele Frequency) - częstotliwość występowania rzadkiego allelu
11. **MF** (Metafounders) - tzw. kody metafounders
12. **MME** (Mixed Model Equations) - układ równań modelu mieszanego
13. **PCG** (Preconditioned Conjugate Gradient) - prekodycjonowana metoda gradientów sprzężonych
14. **RP** (Raw Pedigree) - rodowód, w którym braki danych pozostawione jako kod brakujących danych
15. **SNP** (Single Nucleotide Polymorphism) - polimorfizm pojedycznego nukloetydu
16. **SNP-BLUP** (Single Nucloetide Polymorphism BLUP) - najlepsza liniowa nieobciążona predykcja oparta na markerach polimorfizmu pojedynczego nukleotydu
17. $\boldsymbol{G^{+}P^{+}}$ - osobniki posiadające informacje o genotypie i fenotypie
18. $\boldsymbol{G^{-}P^{+}}$ - osobniki posiadające informacje tylko o fenotypie
19. $\boldsymbol{G^{+}P^{-}}$ - osobniki posiadające informacje tylko o genotypie
20. $\boldsymbol{G^{-}P^{-}}$ - osobniki nie posiadające informacji o genotypie i fenotypie
21. **P_Real** - rodowód pochodzący z rutynowej oceny wartości hodowlanej
22. **P_2010** - rodowód pochodzący z rutynowej oceny wartości hodowlanej ze zwiększoną ilością brakujących danych (~20% dla krów i ~10% dla buhajów)

23. **P_4020** - rodowód pochodzący z rutynowej oceny wartości hodowlanej ze zwiększoną ilością brakujących danych (~40% dla krów i ~20% dla buhajów)

Streszczenie polskojęzyczne

Modele jednostopniowe stają się standardową procedurą używaną w ocenie wartości hodowlanej bydła w wielu krajach. Modele te, cechuje wykorzystanie wszystkich dostępnych informacji o zwierzęciu, tj. informacji fenotypowej, genotypowej i rodowodowej. Korzyść płynąca z użycia modeli jednostopniowych to wspólna ocena wartości hodowlanej zarówno dla zwierząt zgenotypowanych i niezgenotypowanych. Jednakże, złożoność modelu wynikająca z dużej liczby skorelowanych efektów, prowadzi do wyzwań obliczeniowych. Najczęściej stosowane formuły statystyczne to model jednostopniowy genomowej najlepszej liniowej nieobciążonej predykcji (**G-BLUP**) oraz model jednostopniowy najlepszej liniowej nieobciążonej predykcji opartej na markerach polimorfizmu pojedynczego nukleotydu (**SNP-BLUP**).

Celem niniejszej rozprawy było zastosowanie i sprawdzenie jakości predykcji wartości hodowlanych, przy użyciu modeli jednostopniowych w oparciu o populacje bydła Holsztyńsko-fryzyjskiego pochodzącej z rutynowej oceny wartości hodowlanej w Polsce. W szczególności skupiono się na aspektach metodycznych, takich jak: wpływ głębokości rodowodu na ogólną zbieżność modelu, porównaniu różnych podejść modeli jednostopniowych, porównaniu wyników pomiędzy dwoma oprogramowaniami MiXBLUP i BLUPF90 oraz porównaniu różnych podejść kodowania brakujących informacji rodowodowych w zależności od ilości brakujących danych.

Podsumowując, głębokość rodowodu ma znaczący wpływa na tempo zbieżności, im większa liczba pokoleń, tym więcej czasu potrzebuje model do uzyskania zbieżności. Różne podejścia modeli jednostopniowych dają zbliżone wyniki i nie zaobserwowano znaczących różnic, jednakże należy zwrócić uwagę na liczbę osobników rdzeniowych w podejściu **APY** (model **G-BLUP** z użyciem zestawu osobników rdzeniowych), ponieważ zbyt mała liczba może skutkować niedoszacowanymi wynikami. Kolejnym aspektem jest wydajność obliczeniowa, która jest istotna przy analizowaniu dużej liczby danych. Model jednostopniowy **SNP-BLUP** potrzebował najmniej czasu do osiągnięcia zbieżności i zużył najmniej pamięci podręcznej. Porównanie wyników wartości hodowlanych modeli jednostopniowych pomiędzy oprogramowaniami MiXBLUP i BLUPF90 wykazały zbliżone rezultaty. Kodowanie brakujących informacji rodowodowych ma wpływ na jakość oszacowań wartości hodowlanych. Podejścia z użyciem grup genetycznych i kodów metafounders, wykazują lepsze wyniki niż użycie rodowodu z brakami. Jednakże, podejście z użyciem kodów metafounders

może prowadzić do niedoszacowań wartości hodowlanych przy wysokiej niekompletności rodowodu.

Uzyskanie wyniki poszerzają wiedzę na temat modeli jednostopniowych i ich zastosowaniu w ocenie wartości hodowlanej bydła. Dane genomowe są uzupełnieniem informacji wykorzystywanych w konwencjonalnej ocenie wartości hodowlanej, poprawiają zbieżność i dzięki nim model jest odporny na występujące braki danych. Wybór modelu, oprogramowania, podejścia kodowania brakujących rodziców zależy od kompletności, jakości i typu danych.

Streszczenie anglojęzyczne

The single-step model will soon become the standard procedure of most national genetic evaluations of dairy cattle. The use of all available information about the animal, i.e., phenotype, genotype, and pedigree information, characterizes these models. The benefit of using a single-step model is the joint estimation of breeding values for both ungenotyped and genotyped animals. However, the complexity of the model due to the large number of correlated effects leads to computational challenges. The most commonly used models are the single-step genomic best linear unbiased prediction (**G-BLUP**) model and the single-step best linear unbiased prediction based on single nucleotide polymorphism (**SNP-BLUP**).

The purpose of this dissertation was to apply and test the quality of breeding value prediction using single-step models based on the *Holstein-Friesian* cattle population from routine evaluation in Poland. In particular, it focused on methodological aspects such as the effect of pedigree depth on model convergence, comparison of different single-step approaches, comparison between the two programs MiXBLUP and BLUPF90, and comparison of different approaches for coding missing information in the pedigree depending on the amount of missing data.

In summary, the depth of the pedigree has a significant impact on the rate of convergence; the higher the number of generations, the more time the model takes to converge. Different single-step model approaches give similar results, and no significant differences were observed; however, the number of core animals in the **APY** (Algorithm of Proven and Young) approaches has an impact because too few animals as core animals can result in underestimated results. Another aspect is computational efficiency, which is important when we are analyzing large amounts of data. The **SNP-BLUP** single-step model took the least amount of time to converge and used the least amount of memory. Comparisons of breeding values results between MiXBLUP and BLUPF90 software showed similar results. Coding missing pedigree information affects the quality of breeding value estimates. Approaches using genetic groups and metafounders show better results than using a raw pedigree. However, approaches using metafounders can lead to underestimates of breeding values when pedigrees are highly incomplete.

The obtained results expand the knowledge of single-step models and their application in the assessment of the breeding value of cattle. Genomic data complements the information used in conventional breeding value evaluation, improves convergence, and makes the model

resistant to missing data. The choice of model, software, and coding approach for missing parents depends on the completeness, quality, and type of data.

1. Wstęp

Hodowla bydła odgrywa kluczową rolę w produkcji żywności na świecie. Wraz z biegiem czasu i dzięki doskonaleniu metod selekcji, aktualnym trendem selekcyjnym jest utrzymanie wysokiej produktywności stada przy zachowaniu dobrostanu krów. Już w średniowieczu pierwsi hodowcy prowadzili hodowlę za pomocą obserwacji w myśl: „podobne rodzi podobne", gdzie spodziewali się wysokiej mleczności u córek krów, które tę wysoką mleczność miały. Kluczowym pojęciem funkcjonującym w przemyśle hodowlanym jest wartość hodowlana, która jest sumaryczną wartością addytywną efektów wszystkich genów warunkujących daną cechę. Dzięki niej możliwe jest ustalenie rankingu zwierząt hodowlanych oraz wybraniu najlepszych osobników do reprodukcji w celu zagwarantowania postępu genetycznego. Definicja ta funkcjonuje po dziś dzień, jednak metody na podstawie których te wartości zostają obliczone są stale udoskonalane. Na początku XX w. tworzone były księgi hodowlane, a selekcja oparta była na fenotypie i rodowodzie. Następnie w latach 1950-1960, wprowadzona została ocena wartości hodowlanych buhajów na podstawie wydajności córek. Lata 1960-1970 przyniosły rozwój liniowych modeli **BLUP** (najlepsza liniowa nieobciążona predykcja), które polegały na rozwiązaniu układu równań modeli mieszanych zgodnie z pracą Hendersona (1973). Kolejnym przełomem (1990-2000) było bezpośrednie wykorzystanie wydajności poszczególnych próbnych udojów używając regresji losowych i modelowaniu krzywych laktacji, bez konieczności aproksymacji 305 dniowej wydajności laktacyjnej (Ptak i in., 2015). W roku 2008 rozpoczęto w Polsce prace nad rozwojem oceny genomowej, która umożliwiała ocenę młodych buhajów przed uzyskaniem wydajności ich córek, co pozwoliło na przyspieszenie selekcji (Szyda i in., 2011). Natomiast od 2014 roku, zaproponowano kumulacje wszystkich dostępnych źródeł informacji dostępnych dla osobnika (Liu i in., 2014) we wspólnym modelu jednostopniowym, który jest przedmiotem analizy w niniejszej pracy.

Modele jednostopniowe stają się standardową procedurą w rutynowej ocenie bydła w wielu krajach (Legarra i in., 2014; Mantysaari i in., 2017). Zaletą tej metody w porównaniu z metodą dwustopniową jest możliwość zintegrowania wszystkich dostępnych informacji o zwierzęciu tj. danych fenotypowych, genotypowych i rodowodowych. Co więcej, ponieważ jesteśmy w stanie oszacować wartości hodowlane dla wszystkich osobników zarówno zgenotypowanych, jak i niezgenotypowanych, nie ma konieczności publikowaniu dwóch różnych rankingów osobników. Informacje genotypowe są wyrażane przy pomocy polimorfizmów pojedynczych nukleotydów (**SNP**). W badaniach nad implementacjami

rutynowymi wykorzystywane są dwa podejścia jednostopniowe, a mianowicie: model jednostopniowy genomowej najlepszej liniowej nieobciążonej predykcji (**G-BLUP**) (Legarra i in., 2014; Aguilar i in., 2010; Christensen i Lund, 2010) oraz model jednostopniowy najlepszej liniowej nieobciążonej predykcji opartej na markerach polimorfizmu pojedynczego nukleotydu (**SNP-BLUP**) (Liu i in., 2014). Różnica pomiędzy tymi modelami polega na sformułowaniu addytywnej kowariancji genetycznej pomiędzy osobnikami. W modelu jednostopniowym **G-BLUP** macierz kowariancji wartości hodowlanych wszystkich ocenianych osobników obejmuje komponenty związane z osobnikami niezgenotypowanymi, obliczone na podstawie relacji rodowodowych oraz osobników zgenotypowanych, obliczone na podstawie ważonej sumy relacji rodowodowych i informacji genotypowej. Natomiast w modelu jednostopniowym **SNP-BLUP** związek genomowy między addytywnymi efektami genetycznymi **SNP** jest uwzględniony jako oddzielny komponent modelu. Pomimo różnych parametryzacji modele są matematycznie równoważne (Liu i in., 2014; Liu i in., 2015).

Warto pamiętać, że w kontekście krajowej oceny wartości hodowlanej zastosowanie modeli jednostopniowych w dalszym ciągu stanowi wyzwanie zarówno statystyczne, jak i obliczeniowe, ponieważ wymagają one oszacowania kilku milionów efektów, które są ze sobą często silnie skorelowane. Ponadto, kiedy model jest stosowany do dużych zbiorów danych składających się z milionów rekordów, jego rozwiązanie jest obliczeniowo wymagające nie tylko w kontekście zużycia pamięci i wykorzystania procesora, ale również dokładności numerycznej (Cools i in., 2018). Może to stworzyć potencjalne problemy w rozwiązywaniu układu równań modelu mieszanego. W tym celu większość implementacji wykorzystuje prekondycjonowaną metodę gradientów sprzężonych (PCG), która do rozwiązania modeli mieszanych w kontekście analizy dużych danych została zaproponowana przez Strandéna i Lidauera (1999), a następnie była rozwijana przez Vandenplasa i in. (2018, 2019) w kontekście modelu jednostopniowego **SNP-BLUP**.

W zastosowaniu modeli jednostopniowych do oceny rutynowej wykorzystywane są trzy podejścia:

- Oparte na modelu jednostopniowym **G-BLUP**:
    - Podejście **APY**, który nie wykorzystuje całej macierzy relacji genomowych, a jedynie jej część, która jest oparta na zestawie tzw. osobników rdzeniowych. Polega on na odwróceniu macierzy genomowej dla osobników rdzeniowych,

a następnie obliczenia kowariancji genomowej dla pozostałych osobników zgenotypowanych (Misztal i in., 2014).

- Podejście **GT** wyraża odwrotność macierzy kowariancji genomowych, jako funkcję macierzy spokrewnienia odpowiadającej osobnikom zgenotypowanym oraz macierzy wystąpień genotypów **SNP**, obliczonej na podstawie wzoru Woodbury'ego (Misztal i in., 2014).

- Oparte na modelu jednostopniowym **SNP-BLUP**, który ze względu na inne niż w **G-BLUP** sformułowanie struktury kowariancji, wymaga tylko odwrotności diagonalnej macierzy kowariancji **SNP** (Liu i in., 2014).

Kolejnym ważnym aspektem w rutynowej ocenie wartości hodowlanej bydła jest struktura danych rodowodowych (Bradford i in., 2019). Braki danych, które mogą znajdować się w rodowodzie są wyzwaniem w modelowaniu. Standardowym podejściem do radzenia sobie z tym problemem jest zastosowanie grup genetycznych (**GG**), które są zdefiniowane na podstawie kraju pochodzenia, płci i roku urodzenia (Westell i in., 1988; Legarra i in., 2007). Grupy genetyczne są powszechnie stosowane, ale mogą prowadzić do niedoszacowań wartości hodowlanych, zwłaszcza w przypadku bardzo niekompletnych rodowodów (Masuda i in., 2021; Himmelbauer i in., 2014). Alternatywą dla grup genetycznych są tzw. kody metafounders (**MF**) (Legarra i in., 2015) polegające na utworzeniu grup brakujących rodziców w oparciu o pokrewieństwo genomowe oszacowane z informacji o **SNP** ich potomków, przy założeniu częstości alleli wynoszącej 0.5. Kudinov i in. (2020) nie zaobserwowali różnic w oszacowaniach wartości hodowlanych pomiędzy użyciem **GG** i **MF**. Jednakże Bradford i in. (2019), Macedo i in. (2020) oraz Himmelbauer i in., (2024) w przypadku niekompletnych rodowodów, wykazali wyższą dokładność oszacowań wartości hodowlanych przy użyciu **MF** w porównaniu z **GG**.

2. Cel badań

Celem rozprawy było zastosowanie modeli jednostopniowych dla dużej populacji bydła *Holsztyńsko-fryzyjskiego* objętej rutynową oceną wartości hodowlanej oraz skupienie się na aspektach metodycznych, takich jak: dokładność przewidywań wartości hodowlanej, różne warianty modeli jednostopniowych, głębokość rodowodu oraz struktura danych.

3. Hipotezy

Modele jednostopniowe zwiększają dokładność predykcji ze względu na wykorzystanie wszystkich dostępnych informacji tj. informacji fenotypowej, genotypowej i rodowodowej. Struktura oraz poprawność tych trzech dostępnych informacji ma kluczowy wpływ na dokładność uzyskanych predykcji. Głębokość rodowodu ma duże znaczenie, ponieważ zaobserwować można zależność, że im więcej istnieje pokoleń, tym więcej czasu potrzebuje model do uzyskania zbieżności w estymacji parametrów. Z drugiej strony, większa ilość pokoleń pozwala na dokładną estymację kowariancji addytywnie genetycznych pomiędzy osobnikami (**H1**). W literaturze zaproponowano kilka wariantów modeli jednostopniowych, które pomimo różnic w modelach statystycznych uzyskują zbliżone względem siebie wyniki predykcji wartości hodowlanych (**H2**). Również, niezależnie od używanego oprogramowania, które jest dedykowane do rutynowych ocen wartości hodowlanych, przewidywania tychże wartości są zbliżone (**H3**). Rodowód jest ważną składową każdego modelu, z uwagi na konieczność wyrażenia kowariancji addytywnie genetycznej pomiędzy osobnikami występującymi w modelu jako efekt losowy. Jednak występujące braki danych rodowodowych, mogą prowadzić do błędnych predykcji. Dlatego różne podejścia do kodowania brakujących obserwacji rodowodowych, mają wpływa na uzyskane wartości hodowlane (**H4**). W związku z powyższym, rozprawa obejmowała weryfikacje następujących hipotez:

- **H1**: Liczba iteracji w jakiej model jednostopniowy uzyska zbieżność zależy od głębokości rodowodu.
- **H2**: Wyniki oszacowań wartości hodowlanych są zbliżone niezależnie od użytego modelu jednostopniowego.
- **H3**: Oszacowania wartości hodowlanych przy użyciu oprogramowania MiXBLUP i BLUPF90 są zbliżone.
- **H4**: Różne podejścia do kodowania brakujących wartości rodowodowych mają wpływ na oszacowania wartości hodowlanych oraz wyniki walidacji.

4. Materiały i metody

4.1 Optymalizacja funkcji prawdopodobieństwa modelu mieszanego (P1)

Grupy cech oceniane w rutynowej ocenie bydła *Holsztyńsko-fryzyjskiego*, to: produkcja, pokrój, zdrowie wymienia, długowieczność, cechy wycieleniowe, płodność, cechy użytkowo-obsługowe i zdrowotność racic (https://interbull.org/ib/cop_chap6). Wysokość w krzyżu, należąca do cech pokroju została wykorzystana jako cecha modelowa ze względu na wysoką odziedziczalność wynoszącą 0.54 (wycena.izoo.krakow.pl/doc/metody_oceny_2024_1_buhaje.pdf) oraz wysoką dokładność pomiarów fenotypowych.

Pierwszym krokiem badawczym było zaimplementowanie jednocechowego jednostopniowego modelu **SNP-BLUP** (Liu i in., 2014) do predykcji wartości hodowlanych wysokości w krzyżu. Celem implementacji było:

1. zbadanie różnic w zbieżności modelu w zależności od liczby pokoleń uwzględnionych w pliku rodowodowym,
2. zbadanie różnic w zbieżności modelu dla poszczególnych grup zwierząt (**P1**).

Dane (Tabela 1) pochodziły z polskiej rutynowej oceny wartości hodowlanych z grudnia 2021 roku i zawierały: 1,098,611 krów z fenotypem (wysokość w krzyżu), 141,397 buhajów z pseudo-fenotypami odpowiadającymi wartościom hodowlanym poddanym deregresji (**DRP**), które pochodzą z międzynarodowej oceny **MACE** prowadzonej przez Interbull (www.interbull.org), 134,960 zgenotypowanych osobników oraz dwa pliki rodowodowe: pierwszy zawierający wszystkie dostępne pokolenia (FULL) tj. 8,451,809 osobników i drugi zredukowany do piątego pokolenia (5GEN) tj. 1,555,995 osobników. Dane genotypowe zostały zskewencjonowane za pomocą macierzy EuroG MD. W modelu wykorzystano 46,118 **SNP** wspólnych dla wszystkich mikromacierzy. W analizowanej próbie danych częstotliwość rzadkiego allelu (**MAF**) nie przekraczała wartości 0.0064 stosowanej w rutynowej genomowej ocenie wartości hodowlanej. Bazując na tym zbiorze, powstały trzy kolejne zbiory z taką samą liczbą osobników, ale z bardziej rygorystycznym kryterium **MAF**:

- **MAF** ≥ 0.01 – 45,537 **SNP**'ów,
- **MAF** ≥ 0.05 – 41,667 **SNP**'ów,
- **MAF** ≥ 0.01 – 37,380 **SNP**'ów.

Tabela 1. Liczba osobników dla poszczególnych typów danych (**P1**, **P2**, **P3**).

Typ danych	Płeć	Liczba zwierząt	Użyte w publikacji
Wysokość w krzyżu (cały zestaw danych)	Krowy z fenotypem	1,098,611	**P1**, **P2**, **P3**
	Buhaje z **MACE DRP**	141,397	
Wysokość w krzyżu (okrojony zestaw danych)	Krowy z fenotypem	1,001,839	**P2**, **P3**
	Buhaje z **MACE DRP**	134,181	
Kąt racic (cały zestaw danych)	Krowy z fenotypem	1,098,611	**P2**
	Buhaje z **MACE DRP**	141,397	
Kąt racic (okrojony zestaw danych)	Krowy z fenotypem	1,001,839	**P2**
	Buhaje z **MACE DRP**	111,438	
Genotypowe I	Krowy	70,134	**P1**, **P2**
	Buhaje	64,826	
Genotypowe II	Krowy	42,134	**P3**
	Buhaje	47,108	
Rodowodowe			
• Wszystkie pokolenia	Krowy	6,428,481	**P1**
	Buhaje	2,023,328	
• Zredukowany do 5 pokolenia	Krowy	1,368,487	**P1**, **P2**, **P3**
	Buhaje	187,508	

W danych rodowodowych występują brakujące dane, dlatego zostały zaimplementowane grupy genetyczne. Grupy genetyczne służą do zdefiniowania poszczególnych grup osobników z brakującymi rekordami, bazując na ich płci, kraju pochodzenia oraz roku urodzenia. Jak wykazali Tsuruta i in. (2014), Melo i in. (2024) wykorzystanie grup genetycznych pozwala na: zmniejszenie obciążenia przewidywanych wartości hodowlanych, dokładniejszą estymację trendu genetycznego oraz uzyskanie lepszej stabilności obliczeniowej i szybszą zbieżność modelu. Definicje grup genetycznych dla polskiej populacji przedstawiono w Tabeli 2.

Tabela 2. Grupy genetyczne podzielone na kraj pochodzenia, rok urodzenia i płeć (**P1, P2, P3, P4**).

Kraj	Rok urodzenia	Buhaj	Krowa
	< 1960	-99	-99
Polska	1960-1969	-1	-2
Stany Zjednoczone, Kanada	1960-1969	-3	-4
Inne kraje	1960-1969	-5	-6
Polska	1970-1979	-7	-8
Stany Zjednoczone, Kanada	1970-1979	-9	-10
Inne kraje	1970-1979	-11	-12
Polska	1980-1989	-13	-14
Stany Zjednoczone, Kanada	1980-1989	-15	-16
Inne kraje	1980-1989	-17	-18
Polska	1990-1999	-19	-20
Stany Zjednoczone, Kanada	1990-1999	-21	-22
Inne kraje	1990-1999	-23	-24
Polska	2000-2009	-25	-26
Stany Zjednoczone, Kanada	2000-2009	-27	-28
Inne kraje	2000-2009	-29	-30
Polska	2010-2019	-31	-32
Stany Zjednoczone, Kanada	2010-2019	-33	-34
Inne kraje	2010-2019	-35	-36
Polska	2020-obecnie	-37	-38
Stany Zjednoczone, Kanada	2020-obecnie	-39	-40
Inne kraje	2020-obecnie	-41	-42

W celu oszacowania wartości hodowlanych został użyty jednocechowy model jednostopniowy **SNP-BLUP** (Liu i in., 2014):

$$(1)\ \boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{W}_s\boldsymbol{a} + \boldsymbol{e},$$

gdzie $\boldsymbol{y}$ to wektor fenotypów reprezentowanych przez mierzone wartości wysokości w krzyżu krów oraz pseudo-fenotypy w formie **MACE DRP** buhajów, $\boldsymbol{\beta}$ jest to wektor efektów stałych zawierający wiek wycielenia, fazę laktacji i numer obory dla krów oraz sztuczne kody efektów stałych dla buhajów (każdy buhaj miał taki sam kod w obrębie danego efektu stałego), $\boldsymbol{a}$ jest wektorem wartości hodowlanych osobników i grup genetycznych, który jest wyrażony jako $\boldsymbol{a} = \boldsymbol{Zg} + \boldsymbol{u}$, gdzie $\boldsymbol{g}$ jest wektorem losowych efektów **SNP**, $\boldsymbol{u}$ jest losowym wektorem addytywnych efektów poligenicznych, $\boldsymbol{e}$ jest wektorem reszt, $\boldsymbol{X}$, $\boldsymbol{W}_s$ i $\boldsymbol{Z}$ są macierzami wystąpień odpowiednio dla efektów $\boldsymbol{\beta}$, $\boldsymbol{a}$, $\boldsymbol{g}$. Struktura kowariancji modelu jest wyrażona poprzez:

- rozkład wielomianowy normalny $\boldsymbol{g} \sim MNV(\boldsymbol{0}, \boldsymbol{I}\ \frac{1-k}{2\sum_{i=1}^{N} p_i(1-p_i)}\sigma_a^2)$, gdzie stała $k = 0.2$ wyraża proporcje addytywnej wariancji genetycznej do addytywnego efektu poligenicznego. Inaczej mówiąc, część wariancji addytywnie genetycznej niewyjaśnionej przez efekty SNP. p_i jest to frekwencja allelu A w **SNP** i;
- resztowy składnik poligeniczny $\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{A}k\sigma_a^2)$, gdzie $\boldsymbol{A}$ to macierz pokrewieństwa wyestymowana na podstawie rodowodu złożona z następujących komponentów $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix}$, gdzie poszczególne elementy opisują pokrewieństwo pomiędzy osobnikami:
  - $\boldsymbol{A}_{11}$ niezgenotypowanymi,
  - $\boldsymbol{A}_{12}/\boldsymbol{A}_{21}$ zgenotypowanymi i niezgenotypowanymi,
  - $\boldsymbol{A}_{22}$ zgenotypowanymi;
- addytywne wartości hodowlane $\boldsymbol{a} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{H}_s\sigma_a^2)$, gdzie $\boldsymbol{H}_s$ wyrażona jest wzorem $\boldsymbol{H}_s = \begin{bmatrix} \boldsymbol{TGT}^T + \boldsymbol{D} & \boldsymbol{TG} & \boldsymbol{TZB} \\ \boldsymbol{GT}^T & \boldsymbol{G} & \boldsymbol{ZB} \\ \boldsymbol{BZ}^T\boldsymbol{T}^T & \boldsymbol{BZ}^T & \boldsymbol{B} \end{bmatrix}$, gdzie $\boldsymbol{D} = (\boldsymbol{A}_{11})^{-1}, \boldsymbol{T} = \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}, \boldsymbol{B} = \boldsymbol{I}\frac{1}{\sum_{i=1}^{N} 2p_i(1-p_i)}, \boldsymbol{G}$ reprezentuje macierz pokrewieństwa oszacowaną na podstawie genotypu przy użyciu metody VanRaden (2008);
- błąd losowy $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{R}\sigma_e^2)$, gdzie $\boldsymbol{R}$ to macierz diagonalna z wagami dla obserwacji, gdzie waga dla krów wynosi 1, a waga dla buhajów z **MACE DRP** jest funkcją efektywnej liczby córek (**EDC**) (Liu, 2011) dla cechy wysokość w krzyżu.

Addytywna wariancja genetyczna i wariancja resztowa nie były estymowane. Przyjęto wartości parametrów stosowanych w rutynowej ocenie wartości hodowlanych wynoszące $\sigma_a^2 = 5.50$ i $\sigma_e^2 = 4.63$ (prep.interbull.org).

Parametry modelu zostały obliczone przy użyciu oprogramowania MiXBLUP (Vandenplas i in., 2022), w którym do rozwiązania układu równań zaimplementowano dwupoziomową metodę PCG (Vandenplas i in., 2019):

$$(2)\ \boldsymbol{P}^{-1}\boldsymbol{M}^{-1}\boldsymbol{C}\boldsymbol{x} = \boldsymbol{P}^{-1}\boldsymbol{M}^{-1}\boldsymbol{b},$$

gdzie $\boldsymbol{C}$ to macierz współczynników odpowiadająca równaniu modeli mieszanych (**MME**) zdefiniowanego we wzorze (1), $\boldsymbol{x}$ jest wektorem efektów stałych i losowych wyrażonych wzorem $\boldsymbol{x}^T = [\boldsymbol{\beta}^T \boldsymbol{g}^T \boldsymbol{u}^T]$, $\boldsymbol{b}$ to prawa strona równania modelu mieszanego (1), $\boldsymbol{M}$ i $\boldsymbol{P}$ to macierze przedwarunkowe kolejno pierwszego i drugiego poziomu. Współczynnik zbieżności został wyrażony za pomocą kryteriów: CK, CM i CD (Vandenplas i in., 2021). Kryterium CK wyrażone jest wzorem: $CK = \frac{1}{\mu_1}\frac{\|\boldsymbol{M}^{-1}[\boldsymbol{b}-\boldsymbol{C}\hat{\boldsymbol{x}}_i]\|}{\|\hat{\boldsymbol{x}}_i\|}$, gdzie μ_1 jest najmniejszą dodatnią wartością własną macierzy $\boldsymbol{M}^{-1}\boldsymbol{C}$, natomiast i reprezentuje numer iteracji. Kryterium CM wyrażone jest wzorem $CM = k(\boldsymbol{M}^{-1}\boldsymbol{C})\frac{\|\boldsymbol{M}^{-1}[\boldsymbol{b}-\boldsymbol{C}\hat{\boldsymbol{x}}_i]\|}{\|\boldsymbol{M}^{-1}\boldsymbol{b}\|}$, gdzie $k(\boldsymbol{M}^{-1}\boldsymbol{C})$ jest efektywną spektralną liczbą uwarunkowania macierzy $\boldsymbol{M}^{-1}\boldsymbol{C}$. Kryterium CD ma postać $CD = \frac{\|\hat{\boldsymbol{x}}_{i-1}-\hat{\boldsymbol{x}}_i\|}{\|\hat{\boldsymbol{x}}_i\|}$, gdzie wektor $\hat{\boldsymbol{x}}_i$ zawiera estymatory odpowiadające i-tej iteracji. W pracy (**P1**) użyto kryterium zbieżności $CD \leq 1e^{-09}$. Z uwagi na bardzo dużą liczbę iteracji wyniki predykcji wartości hodowlanych oraz wartości kryteriów zbieżności były zapisywane co 20 iteracji. Dodatkowo wzięto pod uwagę wartości bezwzględne pomiędzy ostatecznymi wartościami, a wartościami w poprzednich iteracjach wyrażone wzorem: $|\hat{\boldsymbol{x}}_i - \hat{\boldsymbol{x}}_F|$, gdzie F to estymatory odpowiadające ostatniej iteracji. Różnice były obliczane dla czterech grup zwierząt przedstawionych w Tabeli 3.

Tabela 3. Podział zwierząt ze względu na dostępność danych.

Oznaczenie	Dostępne dane	Liczba osobników
G^+P^+	Genotypowe i fenotypowe	59,242
G^-P^+	Tylko fenotypowe	1,180,846
G^+P^-	Tylko genotypowe	75,718
G^-P^-	Brak genotypu i fenotypu	240,189

Wszystkie obliczenia wykonywano równolegle przy użyciu 12 rdzeni procesora. Dla kontroli niestabilności metody **PCG** związanej z wykonaniem obliczeń równoległych, model z pełną informacją rodowodową został dodatkowo uruchomiony z wykorzystaniem jednego rdzenia.

4.2 Analiza porównawcza wariantów modeli jednostopniowych (P2)

Kolejnym etapem badań było zaimplementowanie różnych wariantów jednocechowych modeli jednostopniowych w obrębie jednego oprogramowania. Celem badania było:

- porównanie algorytmów obliczeniowych pod względem jakości walidacji predykcji wartości hodowlanych dla poszczególnych modeli,
- różnic w oszacowaniach wartości hodowlanych pomiędzy modelami,
- sprawdzenie wydajności obliczeniowej.

Wykorzystano materiał opisany w pracy **P1** (Tabela 1) dla cechy wysokość w krzyżu, wzbogacony o drugą cechę z grup cech pokroju - kąt racicy, charakteryzujący się niską odziedziczalnością h^2=0.09 (wycena.izoo.krakow.pl/doc/metody_oceny_2024_1_buhaje.pdf) o wariancji genetycznej $\sigma_a^2 = 0.10$ (www.interbull.org). Dane dla cechy kąt racicy obejmowały klasyfikację mierzoną w skali od 1 do 9 dostępną dla 1,098,766 krów i pseudo-fenotypy odpowiadające wartościom hodowlanym poddanym deregresji (**DRP**), które pochodzą z międzynarodowej oceny **MACE** prowadzonej przez Interbull (www.interbull.org) dla 117,482 buhajów.

Modele jednostopniowe użyte w porównaniu to wyżej opisany jednocechowy model jednostopniowy **SNP-BLUP** (1) oraz **G-BLUP** (Aguilar i in., 2010), który wyraża się wzorem:

$$(3)\ \boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{W}_G\boldsymbol{a} + \boldsymbol{e},$$

gdzie $\boldsymbol{y}$, to wektor fenotypów reprezentowanych przez mierzone wartości wysokości w krzyżu lub klasyfikacje kąta racic dla krów oraz odpowiadające im pseudo-fenotypy w formie **MACE DRP** buhajów, $\boldsymbol{\beta}$, $\boldsymbol{a}$, $\boldsymbol{e}$ to wektory takie same jak zdefiniowane we wzorze (1). Struktura kowariancji błędu losowego ($\boldsymbol{e}$) jest również taka sama, jak w przypadku wzoru (1), natomiast struktura kowariancji dla wartości hodowlanych jest zdefiniowana jako $\boldsymbol{a} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{H}_G\sigma_a^2)$, gdzie

$\boldsymbol{H}_G = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} + \begin{bmatrix} \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}(\boldsymbol{G}-\boldsymbol{A}_{22})\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} & \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}(\boldsymbol{G}-\boldsymbol{A}_{22}) \\ (\boldsymbol{G}-\boldsymbol{A}_{22})\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} & \boldsymbol{G}-\boldsymbol{A}_{22} \end{bmatrix}$ (Lourenco i in., 2020), gdzie $\boldsymbol{W}_G$ jest macierzą wystąpień dla efektu $\boldsymbol{a}$.

Bardzo duże rozmiary danych odpowiadające całej aktywnej populacji bydła podlegające rutynowej ocenie powodują problemy numeryczne związane z rozwiązywaniem układu równań modelu mieszanego. W pracy **P2** porównano trzy podejścia algorytmiczne do rozwiązania układu równań:

- Podejście **GT** (Mäntysaari i in., 2020), w którym macierz $\boldsymbol{G}^{-1}$ została wyrażona jako $\frac{1}{w}\boldsymbol{A}_{22}^{-1} - \frac{1}{w}\boldsymbol{T}^T\boldsymbol{T}$, gdzie $\boldsymbol{T} = \boldsymbol{L}^{-1}\boldsymbol{Z}^T\boldsymbol{A}_{22}^{-1}$, a w oznacza proporcję resztowej wariancji poligenicznej, $\boldsymbol{L}$ wyraża się wzorem $\boldsymbol{Z}^T\boldsymbol{A}_{22}^{-1}\boldsymbol{Z} + w\boldsymbol{I} = \boldsymbol{L}\boldsymbol{L}^T$.
- Podejście **APY** (Misztal i in., 2014), dzieli zgenotypowane osobniki na dwa podzbiory: zwierzęta rdzeniowe i pozostałe (ang. core, non-core animals), gdzie odwrotność macierzy $\boldsymbol{G}^{-1}$ jest aproksymowana przez $\begin{bmatrix} \boldsymbol{G}_c^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -\boldsymbol{G}_c^{-1}\boldsymbol{G}_{cn} \\ \boldsymbol{I} \end{bmatrix} \boldsymbol{M}_n^{-1}[-\boldsymbol{G}_{nc}\boldsymbol{G}_c^{-1}\ \boldsymbol{I}]$, gdzie c i n oznaczają odpowiednio zwierzęta rdzeniowe i pozostałe, $\boldsymbol{M}_n$ jako macierz diagonalna zawierająca funkcje elementów macierzy $\boldsymbol{G}$, $\boldsymbol{G}_c$ reprezentuje macierz kowariancji addytywnie genetycznych pomiędzy zwierzętami rdzeniowymi, $\boldsymbol{G}_{nc}/\boldsymbol{G}_{cn}$ reprezentuje macierz kowariancji addytywnie genetycznej pomiędzy osobnikami rdzeniowymi i pozostałymi.
- W modelu jednostopniowym **SNP-BLUP** (1) z uwagi na rozdzielenie estymacji wartości hodowlanych i efektów **SNP** nie jest wymagana odwrotności macierzy $\boldsymbol{G}$, a rozwiązanie układu liniowego równań modelu mieszanego wymaga odwrotności diagonalnej macierzy $\boldsymbol{B}$ w $\boldsymbol{H}_s^{-1} = \begin{bmatrix} \boldsymbol{A}^{11} & \boldsymbol{A}^{12} & 0 \\ \boldsymbol{A}^{21} & \boldsymbol{A}^{22} + (\frac{1}{w} - 1)\boldsymbol{A}_{22}^{-1} & -\frac{1}{w}\boldsymbol{A}_{22}^{-1}\boldsymbol{Z} \\ 0 & -\frac{1}{w}\boldsymbol{Z}^T\boldsymbol{A}_{22}^{-1} & \frac{1}{1-w}\boldsymbol{B}^{-1} + \frac{1}{w}\boldsymbol{Z}^T\boldsymbol{A}_{22}^{-1}\boldsymbol{Z} \end{bmatrix}$.

Odpowiednia ilość zwierząt rdzeniowych wpływa na oszacowania wartości hodowlanych, dlatego zastosowano 5 scenariuszy wyboru zgenotypowanych zwierząt do rdzenia:

- **APY3000top** – 3,000 buhajów o najwyższym **EDC**,
- **APY3000random** – 3,000 osobników wybranych losowo,

- **APY10000random** – 10,000 osobników wybranych losowo zgodnie z propozycją Misztal i in. (2015),
- **APY15000top** – 15,000 buhajów o najwyższym **EDC**,
- **APY15000random** – 15,000 osobników wybranych losowo.

4.3 Analiza porównawcza systemów oprogramowania wykorzystywanych do predykcji wartości hodowlanych (P3)

Do rutynowej oceny wartości hodowlanej, używane są dwa oprogramowania MiXBLUP lub BLUPF90 (Aguilar i in., 2018). Głównym celem kolejnej pracy (**P3**) było porównanie oszacowanych wartości hodowlanych wyliczonych za pomocą wyżej wymienionych programów.

Oprogramowanie BLUPF90 korzysta z modelu jednostopniowego **G-BLUP** (3) z podejściem **APY**, dlatego taki model został wykorzystany w obu implementacjach. W celu zapewnienia identycznych danych wejściowych, kryteria zbieżności zostało ustalone na 1e-07 dla MiXBLUP (Vandenplas i in., 2021) oraz 1e-14 dla BLUPF90 (Masuda, 2019), ze względu na różne implementacje techniczne. Do analizy wykorzystano ponownie dane z Tabeli 1 dla cechy wysokość w krzyżu. Do tej pracy przygotowano 4 scenariusze zgenotypowanych zwierząt rdzeniowych:

- **Wszystkie_buhaje** – wszystkie buhaje z pseudo-fenotypami **MACE DRP**,
- **Buhaje_20K** – 20,000 buhajów wybranych losowo,
- **Krowy_30K** – 30,000 krów wybranych losowo,
- **Losowo_20K** – 20,000 osobników wybranych losowo.

Ponadto zostały zdefiniowane trzy scenariusze zwierząt walidacyjnych:

- **Buhaje walidacyjne** - buhaje urodzone w latach 2014-2017 z **EDC** $\geq$ 20,
- **Buhaje z córkami w Polsce** - podzbiór **buhajów walidacyjnych** składający się wyłącznie z buhajów, które mają córki w Polsce,
- **Młode buhaje** - obejmuje młode buhaje z genotypem urodzone od 2018 roku.

4.4 Wpływ brakujących informacji rodowodowych na oszacowania wartości hodowlanych (P4)

Ostatnim etapem było zaimplementowanie najbardziej zaawansowanego modelu - jednostopniowego **SNP-BLUP** z użyciem regresji losowych dla próbnych udojów. Celem tej pracy było porównanie oszacowań wartości hodowlanych przy użyciu różnych scenariuszy kodowania brakujących danych rodowodowych (**P4**). W tej pracy analizowano wyniki walidacji, trendy genetyczne oraz dokładności przewidywań wartości hodowlanych. Dane pochodziły z polskiej rutynowej oceny wartości hodowlanych z kwietnia 2024 (Tabela 4) i zawierały dane: fenotypowe 63,615,019 rekordów próbnych udojów dla wydajności tłuszczu w mleku, genotypowe dla 181,991 osobników z 46,118 **SNP** oraz rodowodowe 4,712,143 zwierząt. Dane rodowodowe zostały uwzględnione do trzeciego pokolenia przy użyciu oprogramowania RelaX2 (Strandén, 2014).

Tabela 4. Liczba osobników dla poszczególnych typów danych (**P4**).

Typ danych	Płeć	Liczba zwierząt	Liczba rekordów
Fenotyp dla wydajności tłuszczu w mleku (pełny zestaw danych)	Krowy	3,707,727	63,615,019
Fenotyp dla wydajności tłuszczu w mleku (okrojony zestaw danych)	Krowy	3,224,917	58,446,695
Genotyp	Krowy	113,019	181,991
	Buhaje	68,972	
Rodowód	Krowy	4,569,044	4,712,143
	Buhaje	143,099	

W oparciu o plik rodowodowy zostały przygotowane trzy scenariusze z różną ilością brakujących danych:

- Oryginalny rodowód (**P_Real**) – rodowód pochodzący z rutynowej oceny wartości hodowlanej zawierający 262,519 (5.6%) brakujących buhajów oraz 719,360 (15.3%) brakujących krów,
- Rodowód 20_10 (**P_2010**) – rodowód pochodzący z rutynowej oceny wartości hodowlanej ze zwiększoną ilością brakujących danych (~20% dla krów i ~10% dla buhajów) zawierający 446,669 (9.5%) brakujących buhajów oraz 1,076,127 (22.8%) brakujących krów,

- Rodowód 40_20 (**P_4020**) – rodowód pochodzący z rutynowej oceny wartości hodowlanej ze zwiększoną ilością brakujących danych (~40% dla krów i ~20% dla buhajów) zawierający 884,192 (18.7%) brakujących buhajów oraz 1,868,957 (39.6%) brakujących krów.

Zmiany w rodowodach zostały wprowadzone dla osobników urodzonych przed 2019 rokiem, a dla młodych osobników rodowód nie został zmieniony. Różnice zaczynają się zatem od drugiego pokolenia, czyli od rodziców młodych osobników.

Dla zdefiniowanych rodowodów wykorzystano trzy sposoby kodowania brakujących rodziców:

- Surowy rodowód (**RP**) – braki danych pozostawione jako kod brakujących danych,
- Grupy genetyczne (**GG**) – brakujące dane reprezentowane przez grupy genetyczne określone na podstawie kraju pochodzenia, płci i roku urodzenia osobnika (Tabela 2),
- Metafounders (**MF**) – brakujące dane zastąpione przez kody metafounders, które reprezentują grupy genetyczne wzbogacone o informacje o genomowym spokrewnieniu osobników.

Liczba grup genetycznych różniła się w zależności od ilości brakujących danych, im więcej braków, tym więcej grup genetycznych. Liczba kodów metafounders jest równa liczbie grup genetycznych, które są podstawą do zastosowania kodów metafounders.

Zastosowano model jednostopniowy **SNP-BLUP** z użyciem regresji losowych dla próbnych udojów do oszacowania wartości hodowlanych (Liu i in., 2004):

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Wf} + \boldsymbol{Vp} + \boldsymbol{Vc} + \boldsymbol{e},$$

gdzie $\boldsymbol{y}$ to wektor fenotypów próbnych udojów dla wydajności tłuszczu pierwszych trzech laktacji każdej krowy, $\boldsymbol{\beta}$ to wektor efektów stałych: stado, dzień doju, częstość doju, $\boldsymbol{f}$ jest wektorem stałych współczynników krzywej laktacji, która jest opisana funkcją Wilmink'a (Liu i in., 2004), $\boldsymbol{p}$ jest wektorem efektów trwałych środowiskowych wyrażonych przez trzy losowe współczynniki wielomianu Legendre'a, $\boldsymbol{c}$ jest losowym wektorem addytywnych efektów poligenicznych również opisanym wielomianem Legendre'a drugiego stopnia, który jest wyrażony jako $\boldsymbol{c} = \boldsymbol{Sq} + \boldsymbol{l}$, gdzie $\boldsymbol{q}$ jest wektorem losowych efektów **SNP**, $\boldsymbol{l}$ jest losowym wektorem addytywnych efektów poligenicznych. $\boldsymbol{X}$ jest macierzą wystąpień efektów stałych $\boldsymbol{\beta}$, $\boldsymbol{W}$ jest macierzą wystąpień dla efektu $\boldsymbol{f}$ oraz $\boldsymbol{S}$ jest macierzą wystąpień dla efektu $\boldsymbol{q}$. $\boldsymbol{V}$ jest macierzą wystąpień dla $\boldsymbol{c}$ oraz $\boldsymbol{p}$ i zawiera wartości dni doju będące funkcją współczynników

wielomianu Legendre'a dla pierwszych trzech laktacji, oraz $\boldsymbol{e}$ jest wektorem reszt. Struktura kowariancji modelu jest wyrażona poprzez rozkład łączony $\boldsymbol{q} \sim N\left(0, (1-k)\boldsymbol{B} \otimes \begin{bmatrix} \boldsymbol{U}_{0_{11}} & \boldsymbol{U}_{0_{12}} & \boldsymbol{U}_{0_{13}} \\ \boldsymbol{U}_{0_{21}} & \boldsymbol{U}_{0_{22}} & \boldsymbol{U}_{0_{23}} \\ \boldsymbol{U}_{0_{31}} & \boldsymbol{U}_{0_{32}} & \boldsymbol{U}_{0_{33}} \end{bmatrix}\right)$, gdzie $\boldsymbol{U}_{0_{ij}}$ to macierz kowariancji genetycznej między laktacjami i i j, która jest zdefiniowana jako $\boldsymbol{U}_{0_{ij}} = \begin{bmatrix} \sigma^2_{11_{ij}} & \sigma^2_{12_{ij}} & \sigma^2_{13_{ij}} \\ \sigma^2_{21_{ij}} & \sigma^2_{22_{ij}} & \sigma^2_{23_{ij}} \\ \sigma^2_{31_{ij}} & \sigma^2_{32_{ij}} & \sigma^2_{33_{ij}} \end{bmatrix}$, gdzie elementy diagonalne to wariancje w danej laktacji, a elementy pozadiagonalne to kowariancje pomiędzy laktacjami, dla genetycznych współczynników regresji losowej. Stała k wyraża proporcje addytywnej wariancji genetycznej do addytywnego efektu poligenicznego, $\otimes$ oznacza iloczyn Kroneckera, a macierz $\boldsymbol{B}$ wyraża się wzorem $\boldsymbol{I}\frac{1}{\sum_{i=1}^{N} 2p_i(1-p_i)}$, addytywne wartości hodowlane $\boldsymbol{l} \sim N\left(0, k\boldsymbol{A}_{22} \otimes \begin{bmatrix} \boldsymbol{U}_{0_{11}} & \boldsymbol{U}_{0_{12}} & \boldsymbol{U}_{0_{13}} \\ \boldsymbol{U}_{0_{21}} & \boldsymbol{U}_{0_{22}} & \boldsymbol{U}_{0_{23}} \\ \boldsymbol{U}_{0_{31}} & \boldsymbol{U}_{0_{32}} & \boldsymbol{U}_{0_{33}} \end{bmatrix}\right)$.

Model został zaimplementowany przy użyciu oprogramowania MiXBLUP.

Sumaryczna wartości hodowlana ($\boldsymbol{GEBV}t$) została obliczona na podstawie predykcji wartości hodowlanych dla trzech pierwszych laktacji ($\boldsymbol{GEBV}1, \boldsymbol{GEBV}2, \boldsymbol{GEBV}3$) według następującego wzoru:

$$\boldsymbol{GEBV}t = 0.5\boldsymbol{GEBV}1 + 0.3\boldsymbol{GEBV}2 + 0.2\boldsymbol{GEBV}3.$$

Walidacja modelu predykcyjnego

Wartości hodowlane przewidziane przez modele opisane w publikacjach (**P2**, **P3**, **P4**) zostały poddane walidacji przy użyciu metody zaproponowanej przez Mäntysaari i in. (2010), w przemyśle hodowlanym określanym jako GEBVtest, który wyraża się wzorem:

$$\boldsymbol{GEBV}f = b_0 + b_1\boldsymbol{GEBV}p + \boldsymbol{e},$$

gdzie $\boldsymbol{GEBV}f$ reprezentuje wektor oszacowanych wartości hodowlanych dla pełnego zbioru danych fenotypowych, $\boldsymbol{GEBV}p$ reprezentuje wektor oszacowanych wartości hodowlanych dla okrojonego zbioru danych fenotypowych, b_0 reprezentuje punkt przecięcia oraz b_1 reprezentuje nachylenie prostej regresji, czyli rozbieżność miedzy wartościami rzeczywistymi reprezentowanymi przez pełny zbiór danych, a przewidywaniami dla okrojonego zbioru

danych. Okrojony zbiór danych został stworzony poprzez usunięcie informacji fenotypowych dla najmłodszych zwierząt (ostatnie 4 lata). Współczynnik R^2, który określa procent wariancji $\boldsymbol{GEBVf}$ wyjaśnionej przez $\boldsymbol{GEBVp}$, został wykorzystany jako miara dokładności przewidywania. Dodatkowo, współczynnik korelacji Pearsona $r(\boldsymbol{GEBVf}, \boldsymbol{GEBVp})$ został użyty jako miara związku liniowego pomiędzy przewidywaniami z pełnego i okrojonego zbioru danych.

Analizy porównawcze oraz wizualizacje zostały przygotowane przy użyciu oprogramowania R (https://www.R-project.org).

5. Wyniki i dyskusja

5.1 Głębokość rodowodu, a liczba iteracji potrzebnych do uzyskania zbieżności modelu (P1)

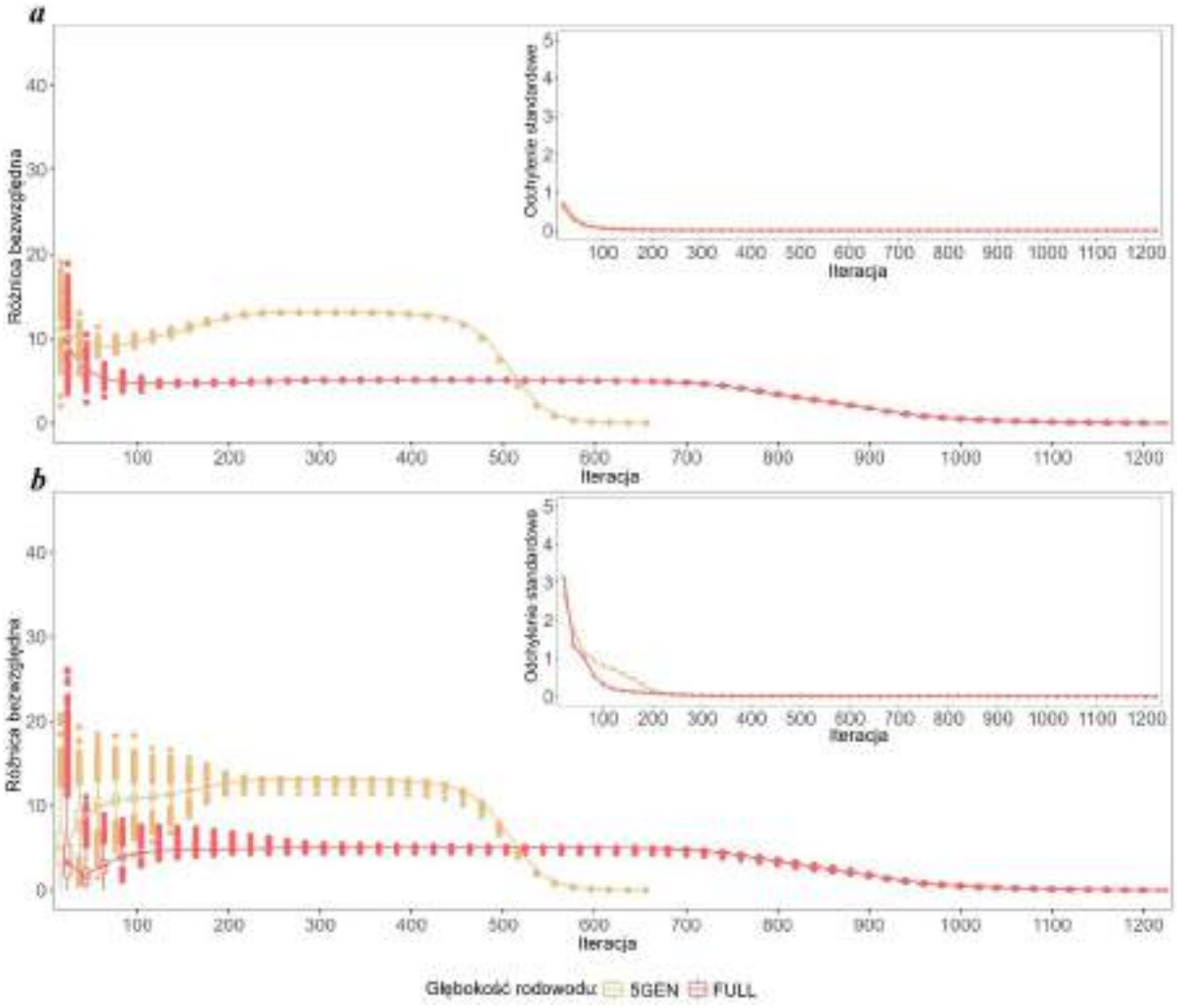
Analiza wzoru zbieżności jednocechowych modeli jednostopniowych uzyskanych w pracy **P1** pokazała, że:

- zbieżność została osiągnięta dwukrotnie szybciej dla zredukowanego zbioru danych,
- wzorce zbieżności wyrażone zmiennością wartości hodowlanych i kryteriów zbieżności w n-tej iteracji z końcowymi wartościami są bardziej widoczne w okrojonym zbiorze danych niż w pełnym zbiorze,
- pełny zbiór danych dawał oszacowania znacznie bardziej zbliżone do ostatecznych rozwiązań niż zredukowany zbiór danych.

Dla zestawów danych z pełnym i okrojonym zestawem danych, wzorzec zbieżności jest podobny niezależnie od różnicy w liczbie iteracji. Po fazie początkowej wskazującej na dużą zmienność oszacowanych wartości hodowlanych (początkowe ~ 200 iteracji), następuje faza stabilizacji charakteryzująca się niewielkimi różnicami w dokładności oszacowań między iteracjami. Ostatnia faza prowadzi do szybkiej (zredukowany zbiór danych) lub monotonicznej (pełny zbiór danych) zbieżności oszacowań efektów modelu.

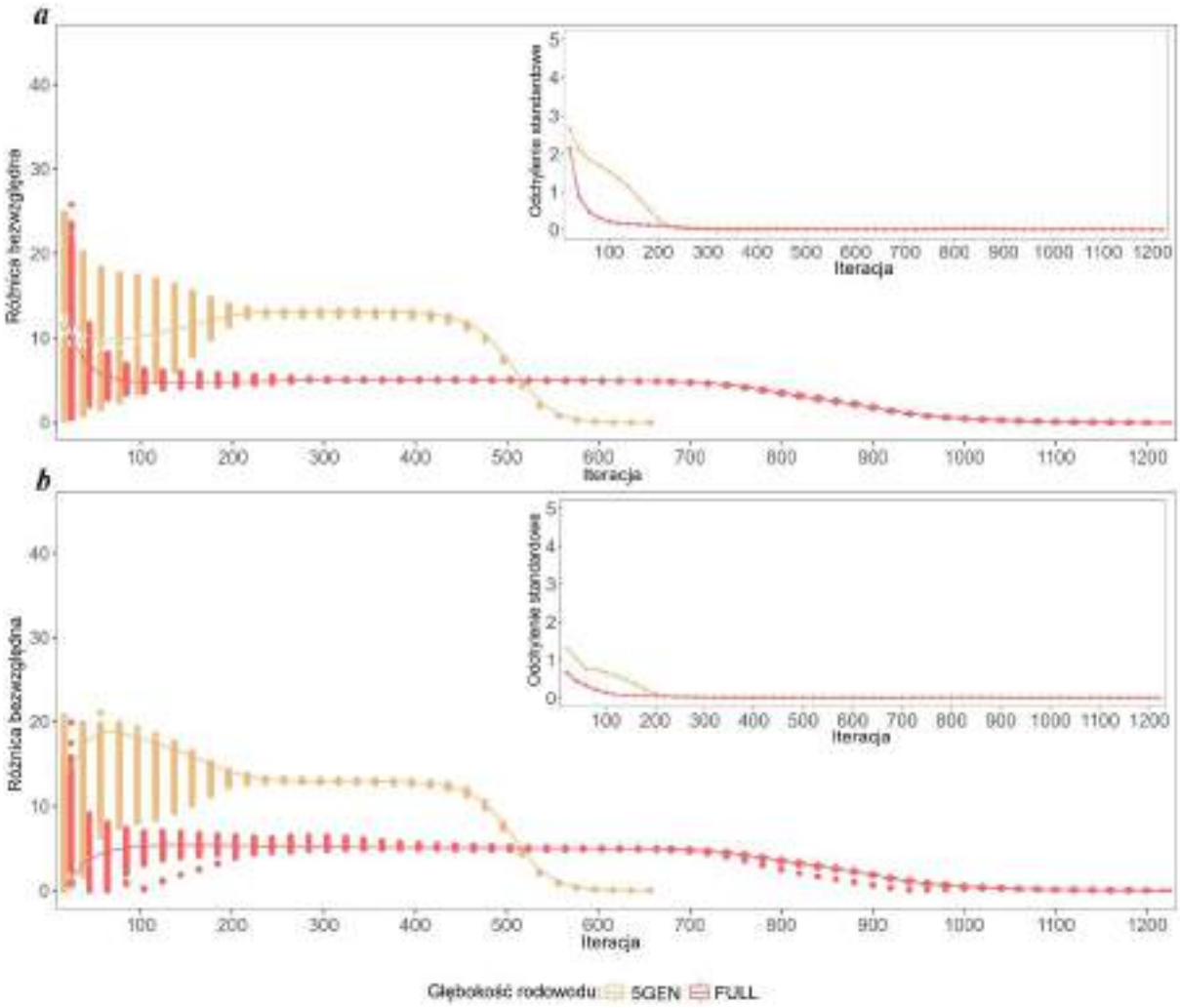
Najmniejsze bezwzględne różnice między ostatecznymi, a pośrednimi oszacowaniami wartości hodowlanych wykazała grupa zwierząt, która posiadała oba źródła informacji G^+P^+, ponadto odchylenia standardowe były najmniejsze wśród wszystkich grup zwierząt (Wykres 1).

Wykres 1. Średnia bezwzględna różnica między ostatecznym, a pośrednimi oszacowaniami wartości hodowlanych oraz ich odchylenia standardowe podczas procesu iteracyjnego dla zwierząt G^+P^+, $\boldsymbol{a}$ reprezentuje krowy, $\boldsymbol{b}$ reprezentuje buhaje.

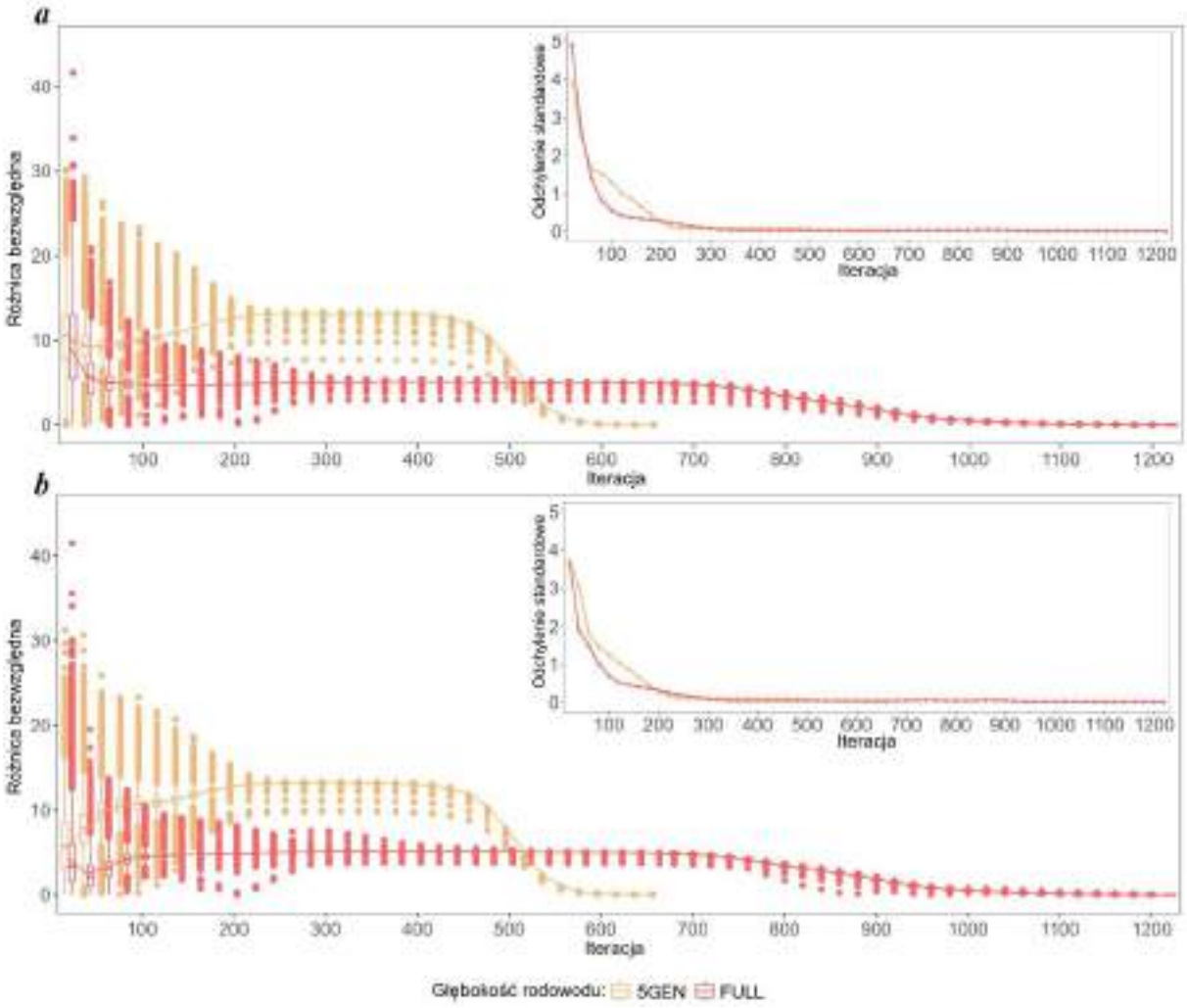


Dla grup G^-P^+ (Wykres 2) i G^+P^- (Wykres 3) trend i bezwzględna średnia różnica oszacowań wartości hodowlanych były podobne jak w przypadku grupy G^+P^+, z tą różnicą, że występowała większa zmienność w przewidywaniach, szczególnie na początku procesu iteracyjnego.

Wykres 2. Średnia bezwzględna różnica między ostatecznym, a pośrednimi oszacowaniami wartości hodowlanych oraz ich odchylenia standardowe, podczas procesu iteracyjnego dla zwierząt $G^{-}P^{+}$, ***a*** reprezentuje krowy, ***b*** reprezentuje buhaje.

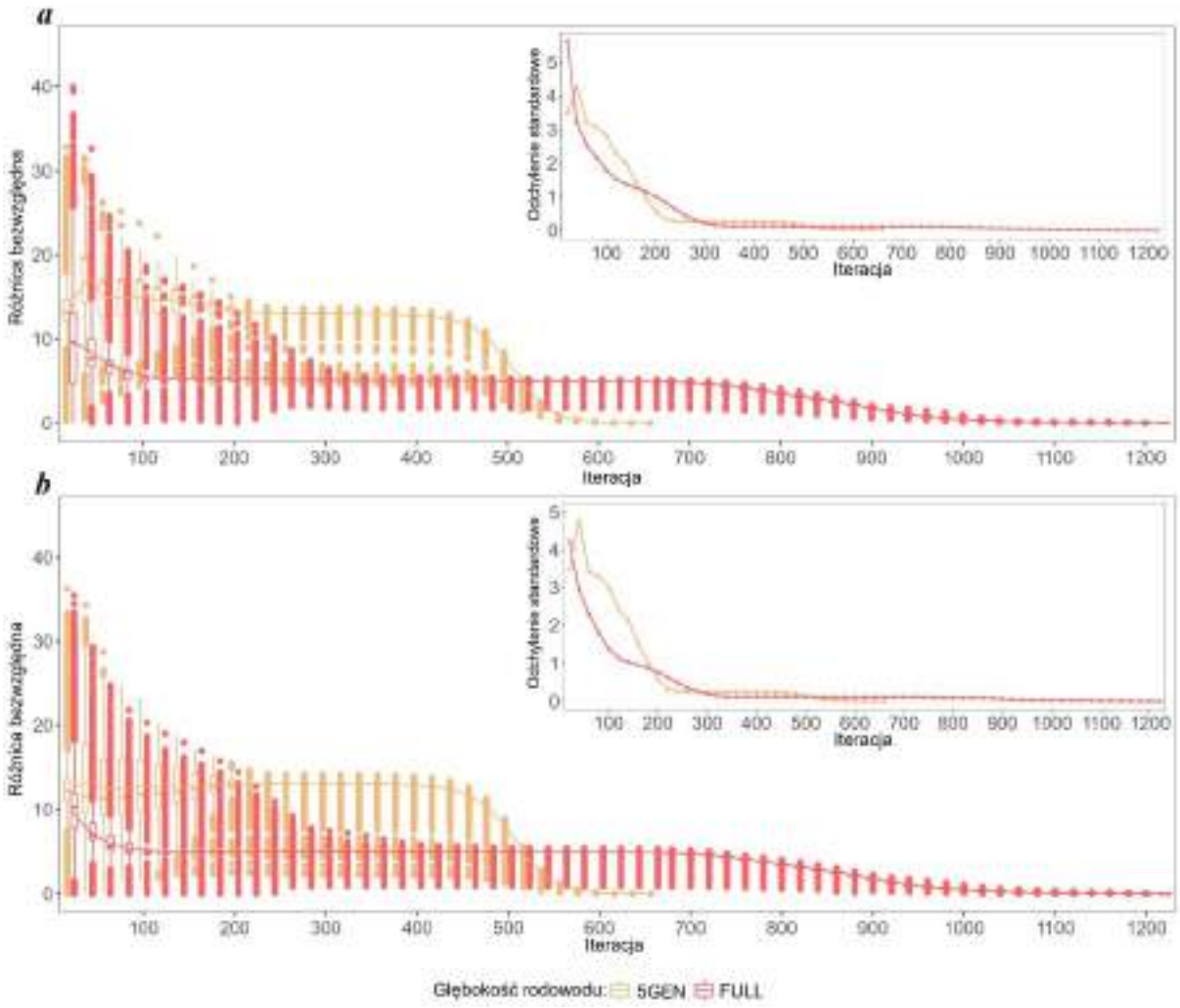


Wykres 3. Średnia bezwzględna różnica między ostatecznym, a pośrednimi oszacowaniami wartości hodowlanych oraz ich odchylenia standardowe, podczas procesu iteracyjnego dla zwierząt $G^{+}P^{-}$, ***a*** reprezentuje krowy, ***b*** reprezentuje buhaje.



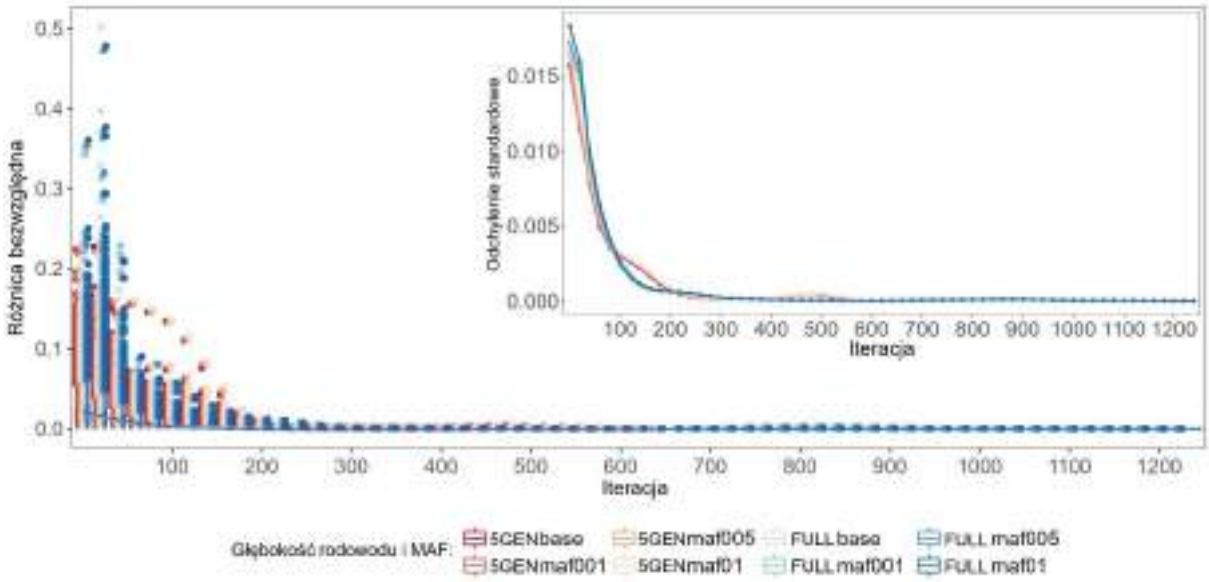
Pomimo podobnego wzorca zbieżności względem pozostałych grup, w grupie zwierząt $G^{-}P^{-}$ zaobserwowano największe rozbieżności względem bezwzględnej średniej różnicy w predykcji wartości hodowlanych podczas procesu iteracyjnego (Wykres 4).

Wykres 4. Średnia bezwzględna różnica między ostatecznym, a pośrednimi oszacowaniami wartości hodowlanych oraz ich odchylenia standardowe, podczas procesu iteracyjnego dla zwierząt $G^{-}P^{-}$, ***a*** reprezentuje krowy, ***b*** reprezentuje buhaje.



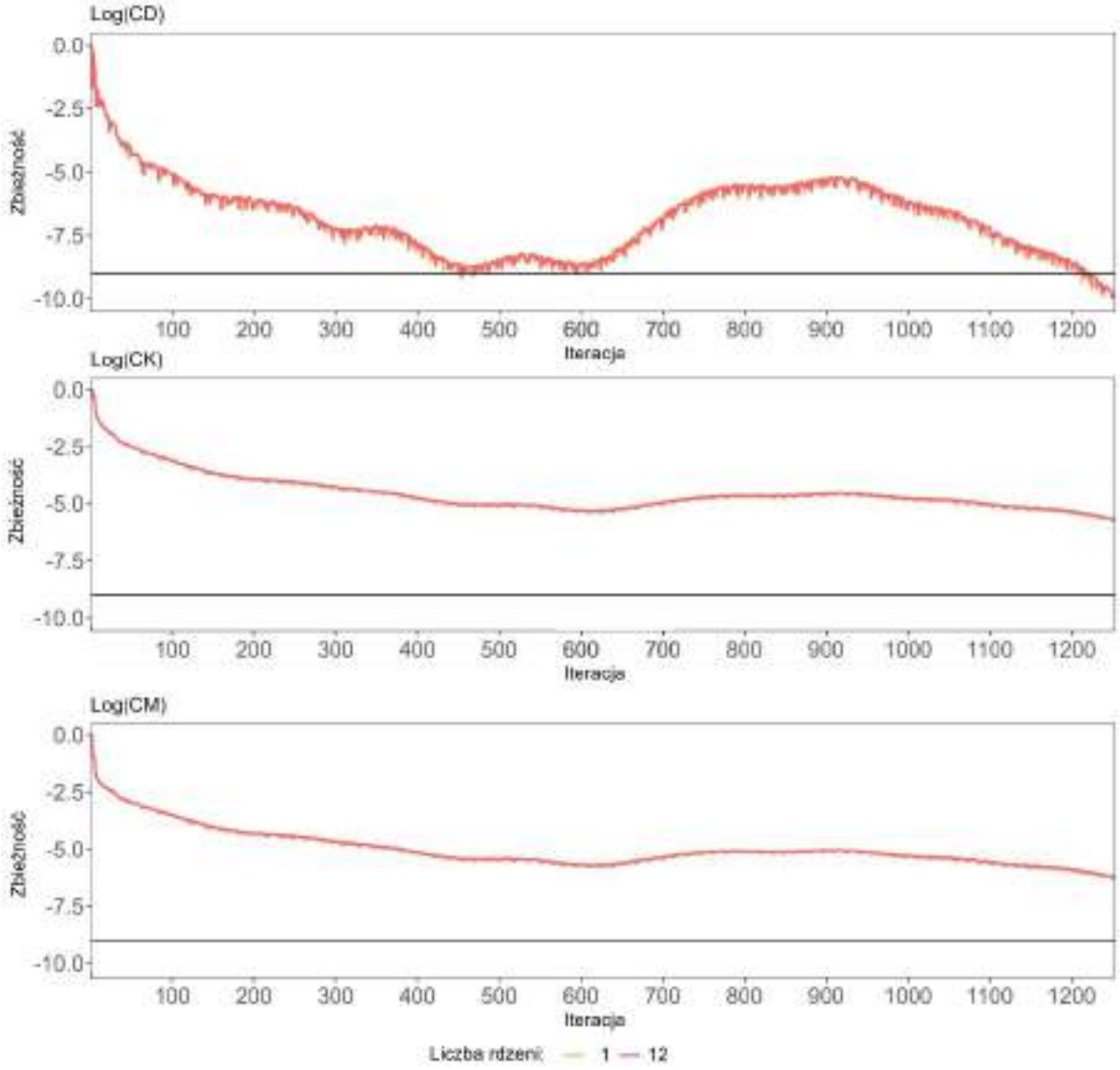
Efekty **SNP** osiągały zbieżność bardzo szybko niezależnie od głębokości rodowodu i częstotliwości rzadkiego allelu. Ostateczne oszacowania zostały osiągnięte w ciągu około 300 iteracji (Wykres 5).

Wykres 5. Średnia bezwzględna różnica w oszacowaniach efektów **SNP** między kolejnymi iteracjami, a ostatecznym wynikiem oraz ich odchylenia standardowe dla różnego progu **MAF**.

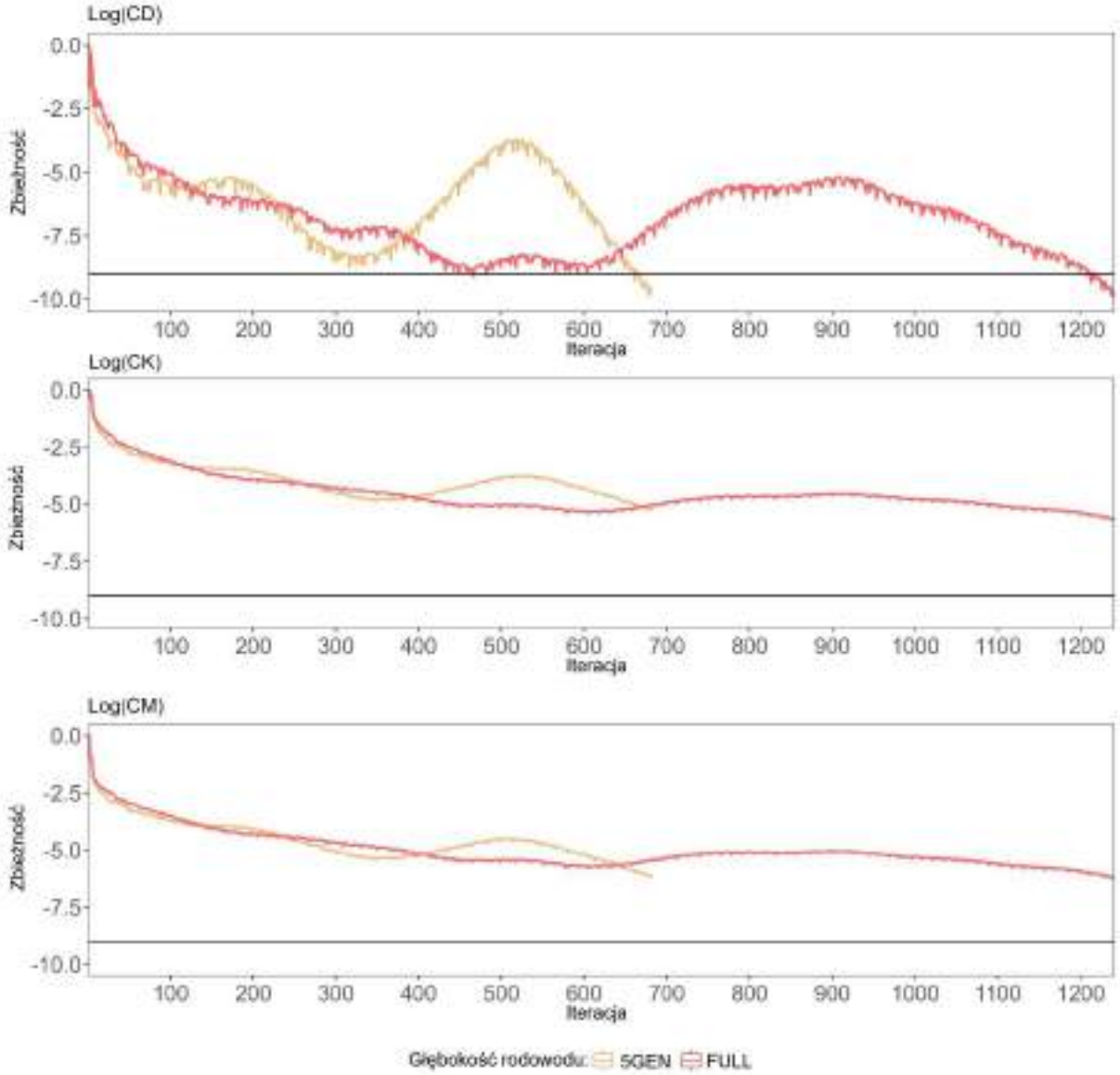


Wyniki procesu iteracyjnego pokazały, że model dla pełnego zestawu danych rodowodowych na 12 rdzeniach osiągnął zbieżność po 1240 iteracjach, natomiast przy użyciu jednego rdzenia liczba iteracji wynosiła 1253 (Wykres 6). Porównując zbieżność pełnego zestawu danych rodowodowych z zestawem zredukowanym do 5 pokolenia, zredukowany zbiór danych osiągnął zbieżność dwukrotnie szybciej (682 iteracje) (Wykres 7). Jednakże oba kryteria zbieżności (CK, CM, CD), wykazały niemonotoniczny spadek przed osiągnieciem ostatecznej zbieżności.

Wykres 6. Kryterium zbieżności (CK, CM, CD) dla jednego i 12 rdzeni dla pełnego zbioru danych. Czarna linia przedstawia kryterium zatrzymania $CD \leq$ 1e-09.



Wykres 7. Kryteria zbieżności (CK, CM, CD) dla pełnego i zredukowanego zbioru danych. Czarna linia przedstawia kryterium zatrzymania $CD \leq$ 1e-09.



Dla pełnego i zredukowanego zestawu danych zaobserwowano wysokie, bliskie 1.0 korelacje Pearsona dla oszacowanych wartości hodowlanych (Tabela 5). Zgodnie z oczekiwaniami najniższe korelacje oszacowano dla grupy z najmniejszą ilością informacji G^-P^- 0.946 (buhaje) oraz 0.980 (krowy).

Tabela 5. Korelacje pomiędzy pełnym, a zredukowanym zestawem danych dla oszacowań wartości hodowlanych dla różnych grup zwierząt.

Grupy zwierząt	Liczba osobników	Korelacja	
		Buhaje	Krowy
Wszystkie osobniki	1,555,995	0.997	0.991
G^+P^+	59,242	0.999	0.999
G^-P^+	1,180,846	0.999	0.999
G^+P^-	75,718	0.999	0.999
G^-P^-	240,189	0.946	0.980

Wyniki oszacowań wartości hodowlanych są bardzo zbliżone pomiędzy pełnym, a zredukowanym zestawem danych przy dwukrotnie szybszej zbieżności procesu iteracyjnego w przypadku zredukowanego rodowodu. Mimo to, rozwiązania dla pełnego zbioru danych z początkowego procesu iteracyjnego były bardziej zbliżone do ostatecznych wyników, co dla jednostopniowego modelu **G-BLUP** wykazał również Pocrnic i in. (2017). Jednakże zaobserwowane przez Legarra i in. (2014) problemy ze zbieżnością mogą powodować mniejszą średnią dokładność uzyskanych oszacowań wartości hodowlanych niezgenotypowanych osobników w przypadku użycia głębokiego rodowodu. Pomimo tego, iż Vandenplas i in. (2021) pokazali w przybliżeniu monotoniczną zbieżność wyrażoną przez kryteria CK, CM, CD, praca ta, jak również inne zastosowania **PCG** w kontekście modeli **SNP-BLUP** i **G-BLUP** (Vandenplas i in., 2018; Pocrnic i in., 2017; Harris i in., 2022) wykazały niemonotoniczny wzorzec zbieżności, charakteryzujący się nie tylko różnym tempem spadku kryterium zbieżności, jak również lokalnym pogorszeniem jakości estymatorów. Ponadto, zaobserwowano również różnice w początkowej szybkości zbieżności pomiędzy poszczególnymi grupami zwierząt ($G^+P^+, G^-P^+, G^+P^-, G^-P^-$). Różnice pomiędzy grupami wynikają ze struktury rodowodu oraz informacji, jakie posiada zwierzę w danej grupie osobników. Stąd oszacowania wartości hodowlanej dla najmniej informatywnej grupy osobników G^-P^- są estymowane na podstawie osobników spokrewnionych, co powoduje większe odchylenia szacowanych wartości hodowlanych podczas procesu iteracyjnego, względem ostatecznych wyników. Interesujący wzorzec zbieżności zaobserwowano podczas porównywania pełnego i zredukowanego zestawu danych rodowodowych. W przypadku pełnego zbioru danych tzw. druga (liniowa) faza zbieżności była znacznie dłuższa, co może wynikać z faktu, że układy równań o dużych wymiarach zazwyczaj nie są tak dobrze

uwarunkowane, jak układy o mniejszych wymiarach, co przekłada się na wydajność procesu iteracyjnego (Pyzara i in., 2011). W przypadku danych wykorzystanych w niniejszej pracy, współczynnik warunkowy, który został obliczony jako iloraz największej i najmniejszej wartości własnej macierzy $\boldsymbol{M^{-1}C}$ był dwukrotnie wyższy dla pełnego (4,258,285), niż dla zredukowanego (2,171,642) zestawu danych, co przekłada się na mniejszy efektywny współczynnik spektralny, który poprawia zbieżność. Ponadto, w przypadku pełnego zbioru danych, większa niestabilność numeryczna wynikająca z zaokrągleń, może pogarszać wydajność zbieżności, ze względu na większa liczbę operacji arytmetycznych (Vandenplas i in., 2018; Cools i in., 2018). Co prawda Cools i in. (2018) wskazali, że zastosowanie obliczeń równoległych powoduje problemy ze zbieżnością, wynikającą z dodatkowej złożoności numerycznej, jednak nie zaobserwowano tego w niniejszej pracy w przypadku użycia jednocechowego modelu jednostopniowego **SNP-BLUP**.

Biorąc pod uwagę wysoką korelację wyników oszacowań wartości hodowlanych pomiędzy pełnym, a zredukowanym zestawem danych, do kolejnych prac wykorzystano zredukowane dane rodowodowe.

5.2 Porównanie oszacowań wartości hodowlanych pomiędzy różnymi modelami jednostopniowymi (P2)

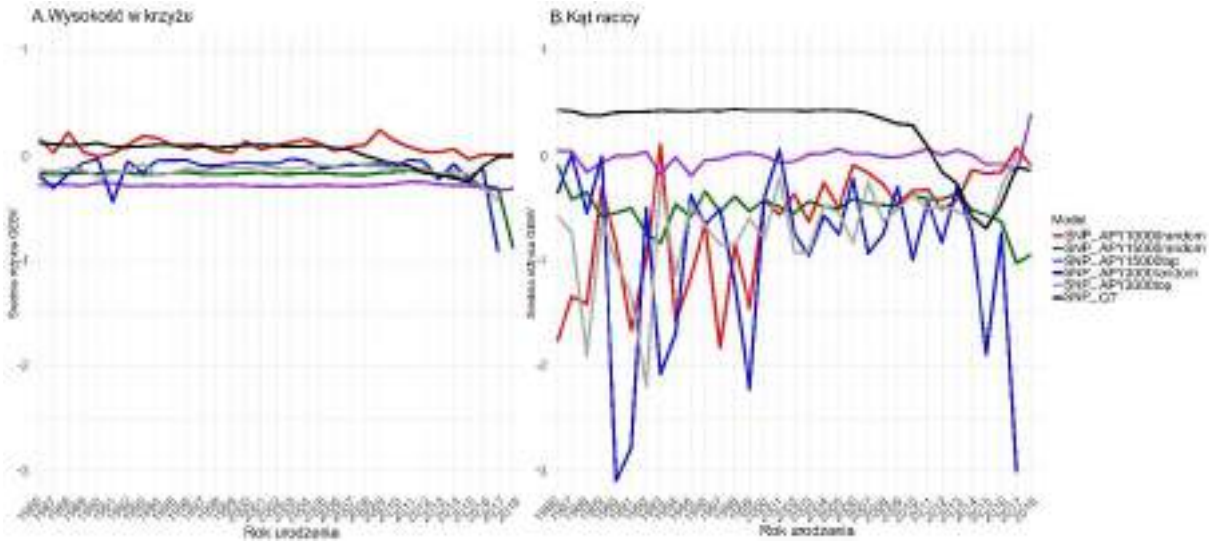
Walidacja modelu to kluczowy etap sprawdzenia czy predykcja wartości hodowlanych jest poprawna. W tym celu przygotowano zbiory walidacyjne dla buhajów z **EDC** $\geq$ 20 zawierające 1,727 buhajów dla cechy wysokość w krzyżu i 1,725 buhajów dla cechy kąt racicy. Wyniki walidacji dla cechy wysokość w krzyżu nie wykazały znaczących różnic w nachyleniu regresji liniowej ($\hat{b}_1$) między poszczególnymi implementacjami modeli jednostopniowych. Estymatory wahały się między 0.94 (**APY3000top**), a 1.02 (**APY10000random**). Co więcej, różnice pomiędzy R^2 były niewielkie z wyjątkiem scenariuszy z 3,000 osobników rdzeniowych, które charakteryzowały się niższymi wartościami R^2, **APY3000top** (0.57) oraz **APY3000random** (0.60). Natomiast podejście **GT** charakteryzowało się najwyższą wartością R^2 wynoszącą 0.82. Podobne wyniki zaobserwowano dla cechy kąt racicy, jednakże wartości R^2 były ogólnie niższe niż dla cechy wysokość w krzyżu. Najwyższe R^2 uzyskano dla podejść **GT** i **SNP-BLUP** (0.76), natomiast najniższe wartości również zaobserwowano dla scenariuszy z 3,000 osobników rdzeniowych, 0.60 (**APY3000random)** oraz 0.62 (**APY3000top**). Szczegółowe wyniki walidacji znajdują się w Tabeli 6.

Tabela 6. Wyniki walidacji dla oszacowań wartości hodowlanych.

Model	$\hat{b}_0$	$\hat{b}_1$	R^2
Wysokość w krzyżu, 1,727 buhajów walidacyjnych, $h^2 = 0.54$			
SNP-BLUP	-3.90 ± 0.32	1.01 ± 0.01	0.77
GT	-4.40 ± 0.27	1.01 ± 0.01	0.83
APY3000top	-1.81 ± 0.44	0.94 ± 0.02	0.57
APY3000random	-2.20 ± 0.44	0.99 ± 0.02	0.60
APY10000random	-3.74 ± 0.36	1.02 ± 0.02	0.72
APY15000top	-2.59 ± 0.32	0.96 ± 0.01	0.75
APY15000random	-2.32 ± 0.33	0.96 ± 0.01	0.73
Kąt racicy, 1,725 buhajów walidacyjnych, $h^2 = 0.09$			
SNP-BLUP	-2.03 ± 0.18	1.03 ± 0.01	0.76
GT	-1.96 ± 0.18	1.04 ± 0.01	0.76
APY3000top	-0.68 ± 0.21	0.97 ± 0.02	0.62
APY3000random	-0.52 ± 0.26	1.03 ± 0.02	0.60
APY10000random	-2.04 ± 0.21	1.02 ± 0.02	0.69
APY15000top	-2.15 ± 0.18	1.02 ± 0.01	0.75
APY15000random	-2.10 ± 0.19	1.03 ± 0.02	0.73

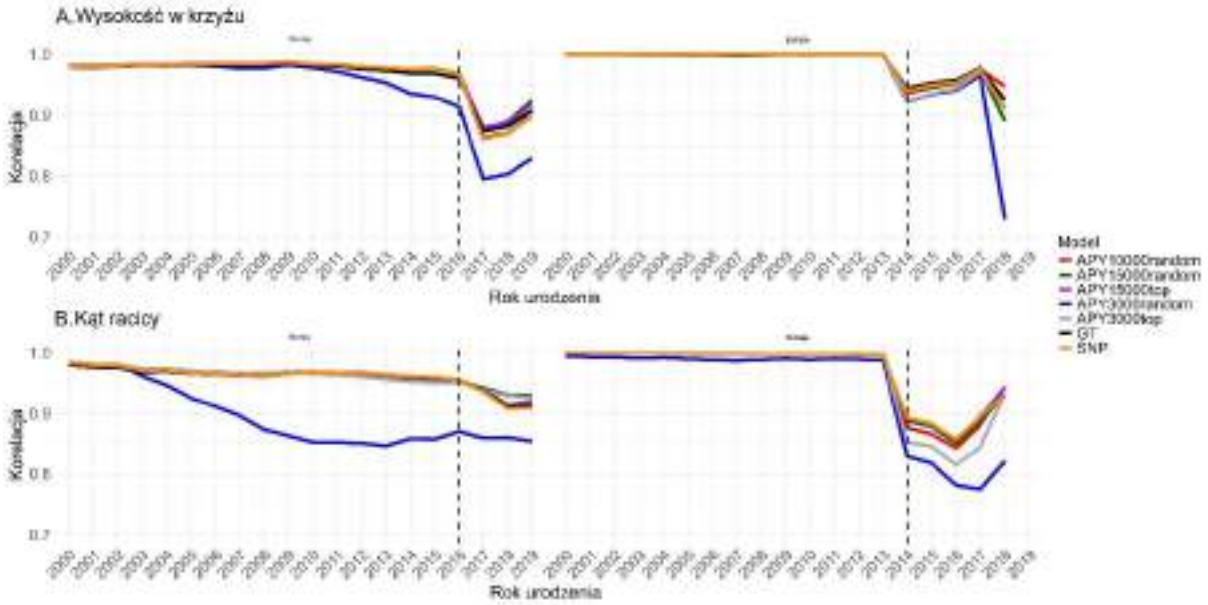
Wykres 8 przedstawia różnice w średnich oszacowaniach wartości hodowlanych pomiędzy podejściem **SNP-BLUP**, a podejściami z użyciem **G-BLUP**. W przypadku porównania **SNP-BLUP** i **GT** uwzględniono oszacowania dla wszystkich zgenotypowanych osobników, natomiast we wszystkich podejściach korzystających z **APY** wykorzystano osobniki rdzeniowe. Oszacowania wartości hodowlanych zostały dodatkowo poddane przeskalowaniu poprzez odjęcie od nich średniej wartości hodowlanej populacji bazowej (krów z fenotypami), charakterystycznym dla każdego z podejść, w celu zapewnienia zbliżonego poziomu. W przypadku wysokości w krzyżu różnice pomiędzy średnimi wartościami oszacowań wartości hodowlanych pomiędzy podejściem **SNP-BLUP**, a podejściami z użyciem **G-BLUP** były niewielkie w poszczególnych latach urodzenia zwierząt (Wykres 8A), natomiast dla kąta racicy wystąpił inny wzorzec (Wykres 8B). Oprócz podejść **GT** oraz **APY15000top**, średnie wartości hodowlane były zróżnicowane w poszczególnych latach urodzenia.

Wykres 8. Różnice pomiędzy średnimi wartościami oszacowań wartości hodowlanych pomiędzy podejściem **SNP-BLUP,** a podejściami z użyciem **G-BLUP**.



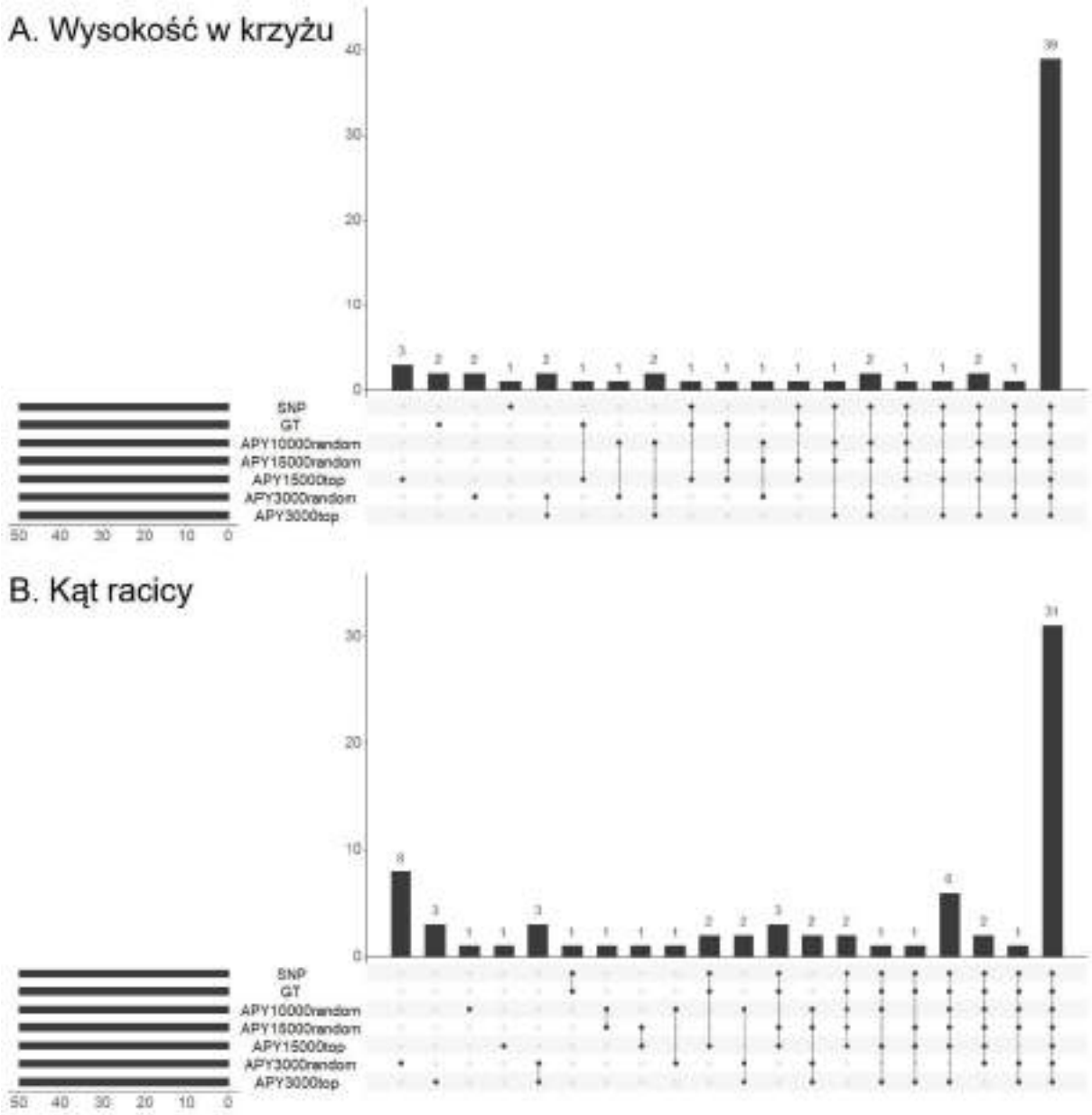
Ponadto, obliczona została korelacja między oszacowaniami wartości hodowlanych uzyskanymi na podstawie pełnego i zredukowanego zestawu danych dla każdego z podejść, podzielonym na krowy (populacja bazowa) oraz buhaje (buhaje, które posiadają co najmniej jedną córkę)(Wykres 9). Linia przerywana dzieli wykres na zwierzęta zdefiniowane jako „stare" i „młode" (te, które zostały usunięte w zredukowanym zestawie danych). W przypadku wysokości w krzyżu niezależnie od modelu, korelacje między pełnym, a zredukowanym zestawem danych dla osobników starych, były zbliżone do 1. W przypadku młodych osobników zaobserwowano tendencję spadkową, przy czym najniższe korelacje obliczono dla **APY3000random** (Wykres 9A). W przypadku kąta racicy uzyskano podobne wyniki, z wyjątkiem **APY3000random** dla krów, gdzie tendencja spadkowa zaczęła się w 2004 roku, a w 2008 roku spadła już poniżej wartości 0.9 (Wykres 9B).

Wykres 9. Korelacja Pearsona pomiędzy pełnym, a zredukowanym zestawem danych dla poszczególnych scenariuszy.



Ostatnim wynikiem, porównującym predykcje w oszacowanych wartościach hodowlanych było stworzenie rankingu 50 buhajów o najwyższych wartościach hodowlanych (Wykres 10). Liczba wspólnych buhajów dla cechy wysokość w krzyżu (Wykres 10A) wyniosła 39, natomiast największą liczbę unikalnych buhajów (3) zaobserwowano w podejściu **APY15000top**. W przypadku kąta racicy (Wykres 10B) liczba wspólnych buhajów była niższa i wynosiła 31. Z kolei aż 8 buhajów było unikalnych przy podejściu **APY3000random**.

Wykres 10. Ranking 50 buhajów o najwyższych wartościach hodowlanych.



Sprawdzono również czas i najwyższą wartość potrzebnej pamięci podręcznej, jaki każdy ze scenariuszy potrzebuje do rozwiązania układu modeli mieszanych. Uogólniając dla obu cech to **SNP-BLUP** oraz podejścia z 3,000 osobników rdzeniowych były najszybsze. Scenariusz **GT** potrzebował najwięcej czasu do otrzymania oszacowań, około 6 razy więcej niż w przypadku **SNP-BLUP**. Najwyższe zużycie pamięci podręcznej dla **SNP-BLUP** było około 10-krotnie niższe, niż w przypadku pozostałych podejść. W kontekście uzyskania zbieżności **SNP-BLUP** wymagał największej liczby iteracji (673 dla wysokości w krzyżu i 1,027 dla kąta racicy) i średnio 2.3 sekundy na iterację. Najmniejszą liczbę iteracji (335) i średnio 0.18

sekundy na iterację dla wysokości w krzyżu potrzebował model **APY3000top**. Szczegółowe podsumowanie zasobów obliczeniowych poszczególnych scenariuszy podano w Tabeli 7.

Tabela 7. Zasoby obliczeniowe, czas oraz liczba iteracji dla poszczególnych scenariuszy.

Model	Czas zegarowy (minuty)		Najwyższa wartość potrzebnej pamięci podręcznej (GB)		Liczba iteracji	
	Wysokość w krzyżu	Kąt racicy	Wysokość w krzyżu	Kąt racicy	Wysokość w krzyżu	Kąt racicy
SNP-BLUP	23	32	5.81	5.81	673	1027
GT	138	143	63.89	63.88	477	629
APY3000top	23	29	49.48	49.47	335	811
APY3000random	23	32	49.48	49.47	390	499
APY10000random	32	34	56.53	56.53	469	652
APY15000top	68	70	61.56	61.56	425	625
APY15000random	54	57	61.56	61.56	477	551

Wiele badań pokazuje przewagę modeli jednostopniowych nad podejściami dwustopniowymi (Alkhoder i in., 2022; Guarini i in., 2018), jednak niewiele skupia się na porównaniu różnych podejść jednostopniowych. Koivula i in. (2012) rozważali modele typu **SNP-BLUP** i **G-BLUP**, jednak skupiali się tylko na zgenotypowanej części populacji. Autorzy zaobserwowali wysokie korelacje oszacowań wartości hodowlanych między poszczególnymi modelami jednostopniowymi, co pokrywa się z wynikami przedstawionymi w niniejszej pracy. Gao i in. (2018) rozważali wydajność obliczeniową walidacji modeli jednostopniowych, i chociaż nie zaobserwowano znaczących różnic to zwrócono uwagę, że w podejściu **APY** liczba oraz zestaw zwierząt rdzeniowych jest kluczowy (Gao i in., 2018; Fragomeni i in., 2015; Masuda i in., 2016; Strandén i in., 2009). Z wyjątkiem podejścia **APY3000random** dla krów, modele **G-BLUP** były odporne na różnice w zestawach zwierząt rdzeniowych. Jednakże, Macedo i in. (2022) wykazali niską odporność modelu **G-BLUP** w kontekście różnych podejść kodowania brakujących rodziców. Ponadto, wyniki niniejszych badań wykazały, że zalecana jest duża ilość osobników rdzeniowych, pod warunkiem dostępności zasobów obliczeniowych, zwłaszcza pamięci podręcznej. Podobnie jak w przypadku wyników Misztal (2015), nie zaobserwowano znaczącej poprawy dokładności oszacowań wartości hodowlanych przy użyciu

więcej niż 10,000 osobników rdzeniowych. Biorąc pod uwagę moc obliczeniową potrzebną do rozwiązania równań modeli mieszanych, zaobserwowano duże różnice w zużyciu pamięci podręcznej między **SNP-BLUP**, a podejściami **G-BLUP**. Podejście **GT** wymagało 10 razy więcej pamięci podręcznej i mniej iteracji niż **SNP-BLUP**, co wykazali również Vandenplas i in. (2023). Jednakże zastosowanie podejścia **GT** wiąże się z wysokimi wymaganiami sprzętowymi i długim czasem obliczeniowym, natomiast **SNP-BLUP** nie zużywa dużej ilości pamięci i jest wydajny obliczeniowo. Ważnym aspektem z punktu widzenia przemysłu hodowlanego jest wynik wykazujący, że oba podejścia nie tworzą identycznych rankingów i nadal istnieje znaczna liczba osobników, które były unikalne w poszczególnych podejściach.

5.3 Porównanie oszacowań wartości hodowlanych pomiędzy systemami oprogramowania (P3)

Porównanie wyników pomiędzy oprogramowaniami MiXBLUP oraz BLUPF90 dało zbliżone wyniki korelacji między oszacowaniami wartości hodowlanych oraz wyniki walidacji. Korelacje wahały się od 0.89 dla scenariusza **Krowy_30K** (**Buhaje z córkami w Polsce**) do 0.98 dla scenariusza **Buhaje_20K** (**Młode buhaje**), niezależnie od oprogramowania. Szczegółowe wyniki korelacji przedstawiono w Tabeli 8.

Tabela 8. Korelacja Pearsona pomiędzy pełnymi, a zredukowanymi zestawami danych dla trzech scenariuszy walidacyjnych.

	Buhaje walidacyjne		Buhaje z córkami w Polsce		Młode buhaje	
	MiXBLUP	BLUPF90	MiXBLUP	BLUPF90	MiXBLUP	BLUPF90
Buhaje_20K	0.92	0.92	0.90	0.90	0.97	0.97
Losowo_20K	0.92	0.92	0.90	0.90	0.96	0.95
Krowy_30K	0.92	0.92	0.90	0.89	0.95	0.94
Wszystkie_ buhaje	0.93	0.93	0.90	0.90	0.96	0.96

Dla każdej z wyróżnionych grup zwierząt walidacyjnych nachylenie prostej regresji ($\hat{b}_1$) oraz współczynnik R^2 były zbliżone pomiędzy oprogramowaniami, natomiast punkt przecięcia ($\hat{b}_0$) różnił się i był wyższy w przypadku oprogramowania BLUPF90. Najlepsze wyniki walidacji otrzymano dla grupy **Młode buhaje**, gdzie $\hat{b}_1$ było bliskie 1, a współczynnik

R^2 wynosił ponad 0.9 i był o około 0.1 wyższy niż w przypadku innych grup. Szczegółowe wyniki walidacji znajdują się w Tabelach 9-11.

Tabela 9. Wyniki walidacji oszacowań wartości hodowlanych dla buhajów urodzonych w latach 2013-2017 z **EDC** ≥ 20.

	$\hat{b}_0$	$\hat{b}_1$	R^2
		MiXBLUP	
Buhaje_20K	4.702	0.844	0.851
Losowo_20K	5.212	0.829	0.849
Krowy_30K	5.442	0.818	0.846
Wszystkie_buhaje	4.608	0.856	0.858
		BLUPF90	
Buhaje_20K	6.954	0.855	0.851
Losowo_20K	5.443	0.845	0.850
Krowy_30K	1.924	0.838	0.831
Wszystkie_buhaje	16.345	0.867	0.860

Tabela 10. Wyniki walidacji oszacowań wartości hodowlanych dla buhajów urodzonych w latach 2013-2017 z **EDC** ≥ 20, których córki urodziły się w Polsce.

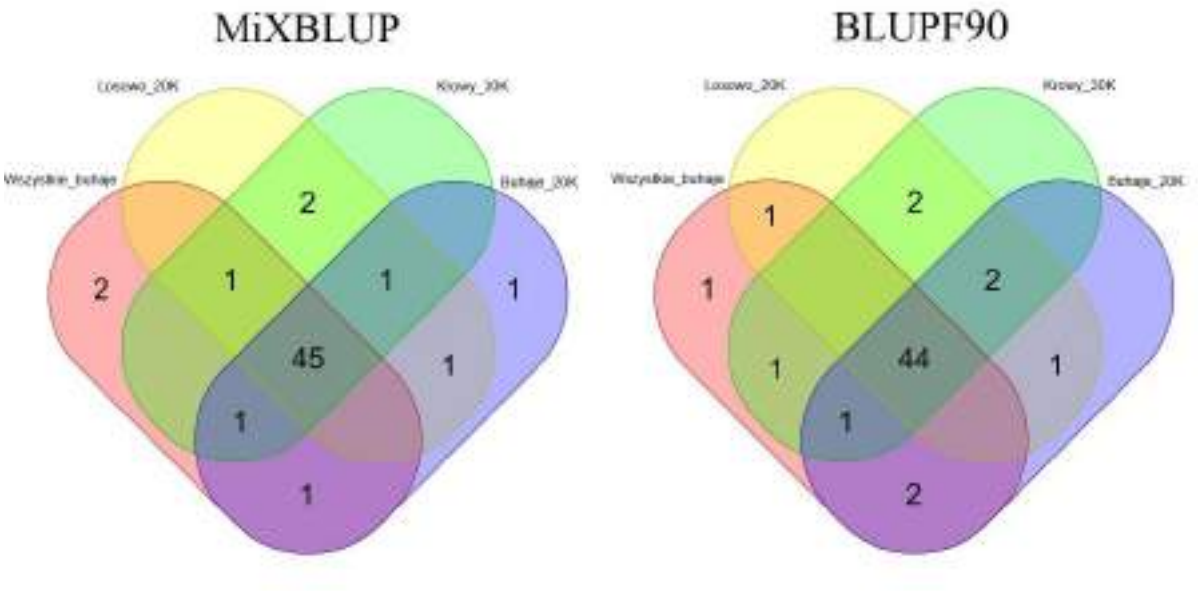
	$\hat{b}_0$	$\hat{b}_1$	R^2
		MiXBLUP	
Buhaje_20K	4.677	0.843	0.805
Losowo_20K	5.175	0.830	0.815
Krowy_30K	5.205	0.830	0.815
Wszystkie_buhaje	4.756	0.848	0.808
		BLUPF90	
Buhaje_20K	6.861	0.861	0.800
Losowo_20K	5.378	0.848	0.814
Krowy_30K	1.615	0.831	0.808
Wszystkie_buhaje	16.504	0.860	0.808

Tabela 11. Wynik walidacji oszacowań wartości hodowlanych dla buhajów z genotypem urodzonych po 2018 roku.

	$\hat{b}_0$	$\hat{b}_1$	R^2
		MiXBLUP	
Buhaje_20K	2.356	0.968	0.938
Losowo_20K	2.273	1.019	0.913
Krowy_30K	1.877	1.030	0.903
Wszystkie_buhaje	3.201	0.959	0.930
		BLUPF90	
Buhaje_20K	4.828	0.980	0.940
Losowo_20K	2.656	1.035	0.910
Krowy_30K	9.839	0.988	0.907
Wszystkie_buhaje	16.175	0.977	0.931

Dalsze porównania pomiędzy oprogramowaniami obejmują ich wysoką odporność na różne scenariusze wyboru zwierząt rdzeniowych (Wykres 11). Ranking 50 najlepszych buhajów pokazuje, że 45 (MiXBLUP) i 44 (BLUPF90) osobniki są wspólne niezależnie od zastosowanego scenariusza wyboru zwierząt rdzeniowych. Porównanie pomiędzy oprogramowaniami wykazało, że wspólnych buhajów jest 48 lub 49 zależnie od scenariusza.

Wykres 11. Ranking 50 najlepszych buhajów dla oszacowań wartości hodowlanych w obrębie każdego oprogramowania.



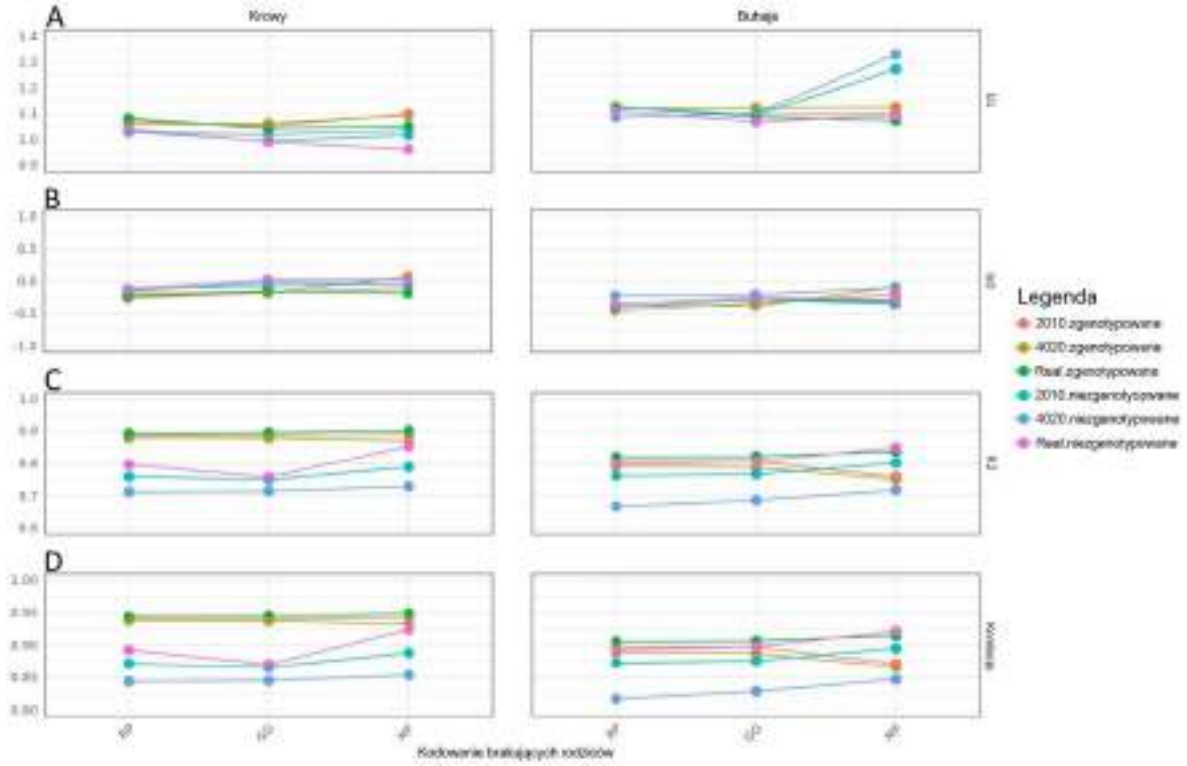
Implementując modele **G-BLUP** w obu oprogramowaniach, MiXBLUP oraz BLUPF90 przy zastosowaniu różnych scenariuszy wyboru zwierząt rdzeniowych uzyskano podobne wyniki w kontekście wydajności obliczeniowej, korelacji pomiędzy oszacowanymi wartościami hodowlanymi i walidacji wartości hodowlanych. Porównanie różnych scenariuszy dotyczących wyboru zwierząt rdzeniowych wykazało wysoką zgodność oszacowanych wartości hodowlanych. Wyniki te są zgodne z założeniami podejścia **APY**, które zakłada, że kluczowy nie jest skład osobników rdzeniowych, a ich liczba (Misztal i in., 2014; Fragomeni i in., 2015). Duża liczba i losowy wybór zwierząt rdzeniowych zapewniają zbliżoną dokładność oszacowanych wartości hodowlanych z zestawem zwierząt wybranymi za pomocą bardziej zaawansowanych kryteriów zaproponowanych przez Misztal i in. (2020). Zbliżone konkluzje uzyskano w niniejszej pracy. Niewielkie różnice występujące pomiędzy tymi dwoma oprogramowaniami, wynikają z implementacji numerycznej algorytmów. Wyniki walidacji przedstawione w niniejszej pracy potwierdzają, że niezależnie od oprogramowania uzyskujemy wiarygodne wyniki oszacowań wartości hodowlanych.

5.4 Porównanie oszacowanych wartości hodowlanych oraz wyników walidacji w różnych implementacjach kodowania brakujących danych rodowodowych (P4)

Przygotowane zostały dwa zbiory walidacyjne: 562 walidacyjnych buhajów (387 zgenotypowanych i 175 niezgenotypowanych) oraz 482,810 walidacyjnych krów (30,227 zgenotypowanych i 452,336 niezgenotypowanych). Wykres 12 przedstawia wyniki walidacji z podziałem na: płeć, status genotypowania, ilość brakujących danych rodowodowych, podejście kodowania brakujących informacji rodowodowych. W przypadku buhajów nachylenia prostych regresji ($\hat{b}_1$) (Wykres 12A) były zbliżone do oczekiwanej wartości 1, z wyjątkiem **P_2010** (1.271) i **P_4020** (1.328) dla niezgenotypowanych osobników z podejściem **MF**. Ponadto, dla podejścia **MF** zaobserwowano większe rozproszenie oszacowań wartości hodowlanych uzyskanych dla **P_2010** i **P_4020** w porównaniu **P_Real**. Wykres 12B przedstawia punkty przecięcia ($\hat{b}_0$), których oszacowania są zbliżone i bliskie 0. Minimalna wartość $\hat{b}_0$wyniosła -0.463 (**P_2010** dla **RP** dla zgenotypowanych buhajów), a maksymalna wartość wyniosła 0.067 (**P_4020** dla **MF** dla zgenotypowanych krów). Współczynniki R^2 (Wykres 13C) i korelacje Pearsona (Wykres 14D) były wyższe dla krów niż dla buhajów, w przypadku zastosowania podejść **MF** lub **GG**. Ponadto, oszacowania wartości hodowlanych zgenotypowanych krów charakteryzowały się wyższym

R^2 i korelacjami niż dla niezgenotypowanych krów, niezależnie od ilości brakujących danych rodowodowych. Nieoczekiwanie, w scenariuszu **P_Real** dla niezgenotypowanych krów najniższą korelację oraz R^2 zaobserwowano dla podejścia **GG** (0.870;0.760) w porównaniu z podejściami **RP** i **MF**. Korelacje dla niezgenotypowanych buhajów wzrastały od podejścia **RP**, przez podejście **GG** do podejścia **MF**. Natomiast w przypadku zgenotypowanych buhajów dla scenariusza **P_Real** korelacja była podobna we wszystkich podejściach, natomiast **P_2010** i **P_4020** dla podejścia **MF** dawały niższe korelacje niż w przypadku podejść **RP** i **GG**.

Wykres 12. Wyniki walidacji z podziałem na: płeć, status genotypowania, ilość brakujących danych rodowodowych oraz podejście kodowania brakujących informacji rodowodowych.

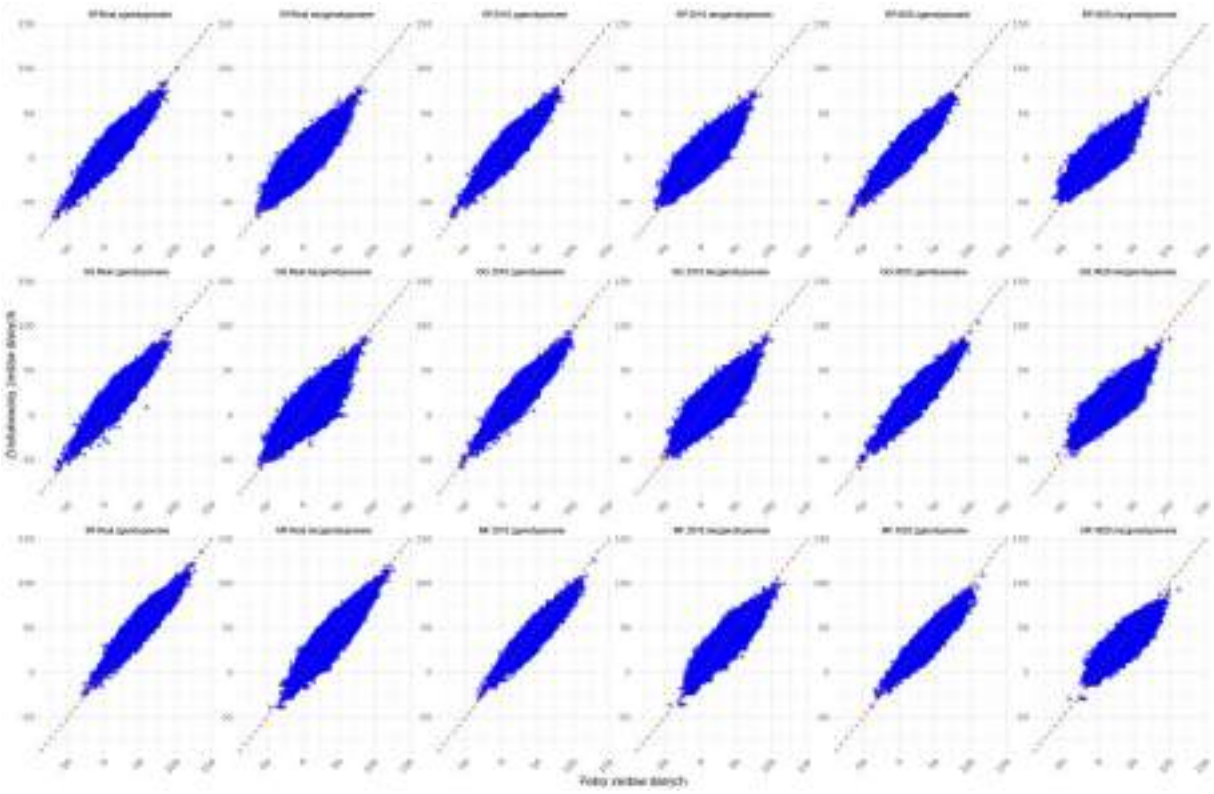


A - nachylenie prostej regresji ($\hat{b}_1$)
B - punkt przecięcia ($\hat{b}_0$)
C - współczynnik R^2
D - korelacja Pearsona

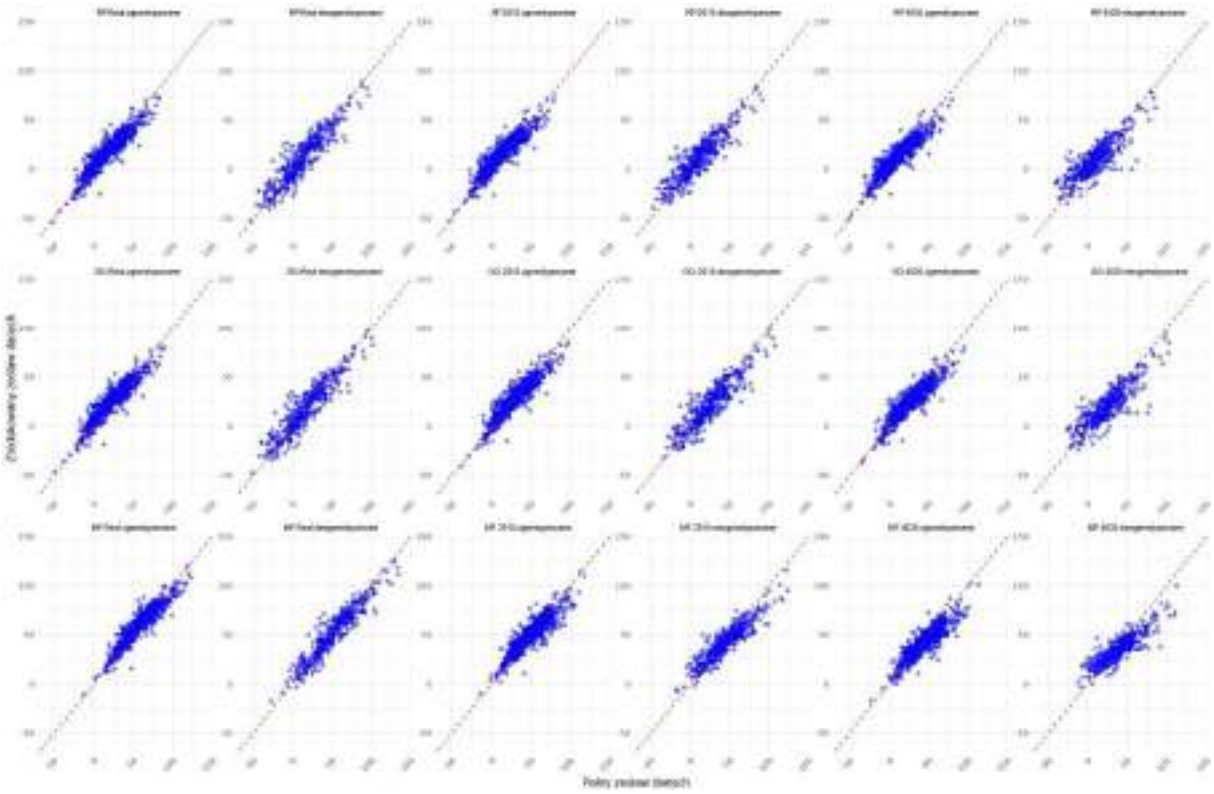
Porównanie oszacowanych wartości hodowlanych dla pełnych i zredukowanych zestawów danych zostało zaprezentowane na Wykresach 13 i 14. W każdym scenariuszu punkty na wykresach układały się w eliptyczny kształt wzdłuż prostej linii regresji. Jednak w przypadku osobników niezgenotypowanych, zwłaszcza dla krów (Wykres 13), zaobserwowano większe rozproszenie wzdłuż przekątnej. Dla buhajów (Wykres 14) wyższe rozproszenie zaobserwowano dla niezgenotypowanych osobników szczególnie dla scenariusza

P_4020, gdzie oszacowania wartości hodowlanych dla zredukowanego zestawu danych były przeszacowane, na co również wskazuje wynik walidacji ($\hat{b}_1 = 1.328$).

Wykres 13. Porównanie oszacowań wartości hodowlanych pomiędzy pełnym, a zredukowanym zestawem danych dla krów z podziałem na status genotypowania, ilość brakujących danych rodowodowych oraz podejście kodowania brakujących informacji rodowodowych.

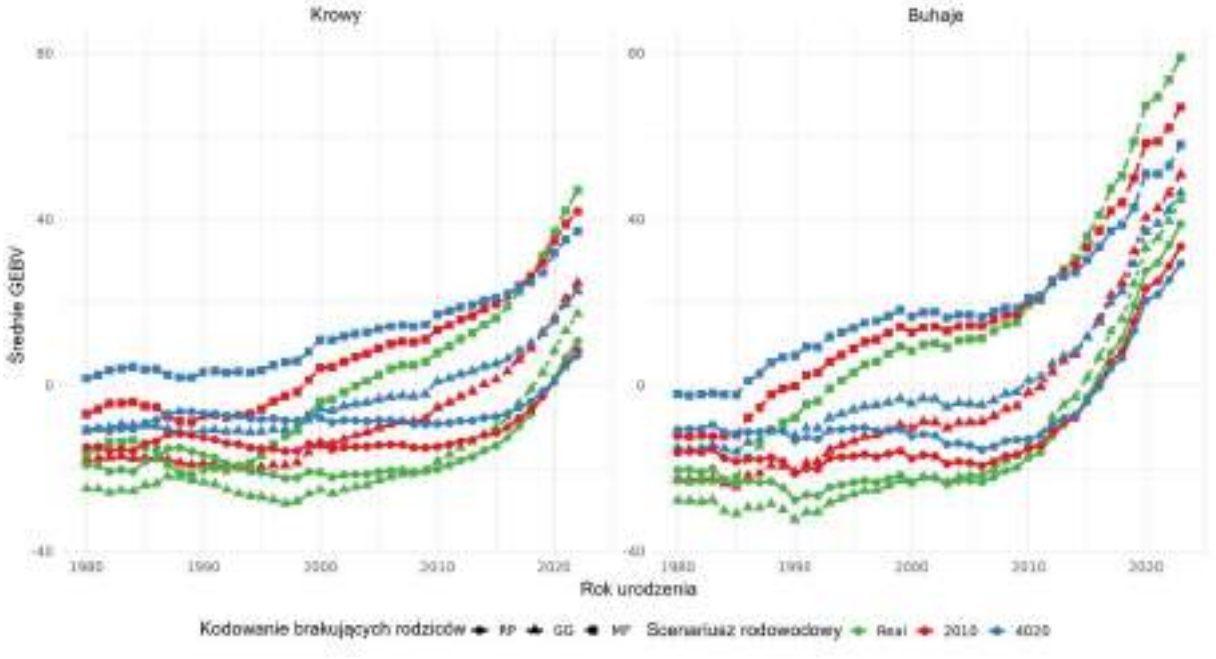


Wykres 14. Porównanie oszacowań wartości hodowlanych pomiędzy pełnym, a zredukowanym zestawem danych dla buhajów z podziałem na status genotypowania, ilość brakujących danych rodowodowych oraz podejście kodowania brakujących informacji rodowodowych.



Wykres 15 przedstawia średni trend oszacowań wartości hodowlanych z podziałem na: płeć, ilość brakujących danych rodowodowych oraz podejście kodowania brakujących informacji rodowodowych w poszczególnych latach urodzenia osobników. Po 2010 roku zaobserwowano wzrost średniej wartości hodowlanej dla wszystkich scenariuszy, zwłaszcza dla buhajów. W każdym ze scenariuszy najwyższy wzrost średniej wartości hodowlanej osiągnął scenariusz **P_Real** z użyciem podejścia **MF**.

Wykres 15. Średni trend oszacowań wartości hodowlanych z podziałem na: płeć, ilość brakujących danych rodowodowych oraz podejście kodowania brakujących informacji rodowodowych.



Badanie skupiało się na porównaniu różnych metod kodowania brakujących rodziców (**RP**, **GG**, **MF**) z wykorzystaniem modelu jednostopniowego **SNP-BLUP** z użyciem regresji losowej dla próbnych udojów. Zbadano, na ile te trzy podejścia są odporne na rosnącą ilość brakujących informacji rodowodowych, osobno dla zgenotypowanych i niezgentoypowanych zwierząt. Dostępność informacji genomowej poprawia jakość oszacowań wartości hodowlanych poprzez uzyskanie lepszego dopasowania do prostej regresji, wyższym współczynnikiem R^2 oraz korelacją Pearsona, pomiędzy pełnym, a zredukowanym zestawem danych w modelu walidacyjnym. Dzięki dostępności informacji genomowej wybór metody kodowania brakujących rodziców nie ma wpływu na jakość oszacowań wartości hodowlanych. W przypadku osobników niezgenotypowanych, podejście **MF** nie jest odporne na zwiększoną ilość brakujących rodziców. Wraz ze wzrostem liczby kodów **MF** wzrastało niedoszacowanie wartości hodowlanych. Szczególnie w przypadku niezgenotypowanych buhajów, poprzez zwiększoną wartość nachylenia prostej regresji ($\hat{b}_1$). W literaturze porównywano głównie metody kodowania brakujących rodziców w kontekście zastosowania modelu jednostopniowego **G-BLUP** (Garcia-Baccino i in., 2017; Macedo i in., (2020, 2022); Kluska i in., 2021). Praca Garcia-Baccino i in. (2017) opierała się na danych symulowanych i wykazała wysoką skuteczność predykcyjną modelu z użyciem **MF**. W przypadku populacji owiec mlecznych Macedo i in. (2020, 2022) uzyskali najdokładniejsze wyniki oszacowań wartości hodowlanych przy użyciu **MF**, wyrażone przez nachylenie prostej regresji bliskie 1. W pracy

Kluska i in. (2021) uzyskano bardzo podobne wyniki przy zastosowaniu **MF** i **GG**. Jednakże podejście **MF** zostało rekomendowane, jako podejście zapewniające najmniejsze obciążenie predykcji.

6. Konkluzje

Wyniki uzyskane w niniejszej pracy pokazały, że w celu uzyskania zbieżności, osobniki bez informacji fenotypowej i genotypowej potrzebują większej liczby iteracji niżeli osobniki, które posiadają zarówno informacje fenotypową, jak i genotypową. Głębokość rodowodu ma znaczący wpływ na tempo zbieżności, przy czym im mniejsza liczba pokoleń w rodowodzie, tym model szybciej uzyska zbieżność. W kontekście predykcji wartości hodowlanych przy użyciu różnych wariantów modeli jednostopniowych, nie zaobserwowano znaczących różnic, z wyjątkiem modelu **APY** z 3,000 osobników rdzeniowych. Jednakże modele **GT** oraz **SNP-BLUP** wykazały najwyższą dokładność oszacowań wartości hodowlanych, podczas gdy dokładność modelu **APY** zależy od wielkości zbioru zwierząt rdzeniowych. W kontekście analizy dużych zbiorów danych odpowiadającym krajowym populacjom bydła mlecznego, ważnym czynnikiem wpływającym na wybór modelu jest wydajność obliczeniowa, gdzie **SNP-BLUP** zużywa najmniej pamięci podręcznej i potrzebuje najmniej czasu do osiągnięcia zbieżności. Natomiast porównanie oszacowań wartości hodowlanych pomiędzy oprogramowaniami MiXBLUP i BLUPF90 wykazało, że implementacje w programach dają zbliżone wyniki, dlatego wybór programu do rutynowej oceny wartości hodowlanej powinien opierać się na innych kryteriach (cena, wsparcie informatyczne, dokumentacja, efektywność obliczeniowa). Różne podejścia kodowania brakujących informacji rodowodowych mają wpływ na jakość oszacowań wartości hodowlanych. Podejścia **MF** i **GG**, wykazują lepsze modelowanie niż użycie surowego rodowodu. Jednakże wraz ze wzrostem ilości brakujących informacji rodowodowych, podejście **MF** może prowadzić do niedoszacowań wartości hodowlanych, dlatego wybór kodowania brakujących rodziców zależy od kompletności danych.

W konkluzjach końcowych pragnę podkreślić, że w analizach przeprowadzonych w niniejszej pracy został wykorzystany rzeczywisty zbiór danych odpowiadający krajowej rutynowej ocenie wartości hodowlanej bydła, dzięki temu wyniki można odnieść do innych krajowych populacji. Dane genomowe, które są ważnym elementem modeli jednostopniowych, są ważnym uzupełnieniem informacji wykorzystywanych w konwencjonalnej ocenie wartości hodowlanej. Poprawiają zbieżność i sprawiają, że model jest odporny na występujące braki danych rodowodowych, co jest kluczową cechą każdego modelu implementowanego w kontekście rutynowym. Jednakże wykorzystanie wszystkich dostępnych źródeł informacji, generuje zwiększone zapotrzebowanie na moc obliczeniową, dlatego wybór odpowiedniego

modelu, oprogramowania i architektury obliczeniowej zależy od wielkości analizowanych populacji i częstotliwości rutynowych ocen.

7. Bibliografia

1. Aguilar, I., Misztal, I., Johnson, D., Legarra, A., Tsuruta, S., & Lawlor, T. (2010). Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *Journal of Dairy Science*, *93*(2), 743–752. https://doi.org/10.3168/jds.2009-2730
2. Aguilar, I., Tsuruta, S., Masuda, Y., Lourenco, D.A.L., Legarra, A., & Misztal, I. (2018). BLUPF90 suite of programs for animal breeding with focus on genomics. In *Proceedings of 11th World Congress on Genetics Applied to Livestock Production*.
3. Alkhoder, H., Liu, Z., Segelke, D., & Reents, R. (2022). Comparison of a single-step with a multistep single nucleotide polymorphism best linear unbiased predictor model for genomic evaluation of conformation traits in German Holsteins. *Journal of Dairy Science*, *105*(4), 3306–3322. https://doi.org/10.3168/jds.2021-21145
4. Bradford, H., Masuda, Y., VanRaden, P., Legarra, A., & Misztal, I. (2019). Modeling missing pedigree in single-step genomic BLUP. *Journal of Dairy Science*, *102*(3), 2336–2346. https://doi.org/10.3168/jds.2018-15434
5. Christensen, O. F., & Lund, M. S. (2010). Genomic prediction when some animals are not genotyped. *Genetics Selection Evolution*, *42*(1), 2. https://doi.org/10.1186/1297-9686-42-2
6. Cools, S., Fatih Yetkin, E., Agullo, E., Giraud, L., & Vanroose, W. (2018). Analyzing the effect of local rounding error propagation on the maximal attainable accuracy of the pipelined conjugate gradient method. *SIAM Journal on Matrix Analysis and Applications*, *39*, 426–450. https://doi.org/10.1137/17M1117872
7. Fragomeni, B., Lourenco, D., Tsuruta, S., Masuda, Y., Aguilar, I., Legarra, A., Lawlor, T., & Misztal, I. (2015). Hot topic: Use of genomic recursions in single-step

genomic best linear unbiased predictor (BLUP) with a large number of genotypes. *Journal of Dairy Science*, *98*(6), 4090–4094. https://doi.org/10.3168/jds.2014-9125

8. Garcia-Baccino, C. A., Legarra, A., Christensen, O. F., Misztal, I., Pocrnic, I., Vitezica, Z. G., & Cantet, R. J. C. (2017). Metafounders are related to F st fixation indices and reduce bias in single-step genomic evaluations. *Genetics Selection Evolution*, *49*(1), 34. https://doi.org/10.1186/s12711-017-0309-2
9. Guarini, A., Lourenco, D., Brito, L., Sargolzaei, M., Baes, C., Miglior, F., Misztal, I., & Schenkel, F. (2018). Comparison of genomic predictions for lowly heritable traits using multi-step and single-step genomic best linear unbiased predictor in Holstein cattle. *Journal of Dairy Science*, *101*(9), 8076–8086. https://doi.org/10.3168/jds.2017-14193
10. Harris, BL., Sherlock, RG., & Nilforooshan, MA. (2022). Large-scale multiple-trait single-step marker model implementation. In *Proceedings of 12th World Congress on Genetics Applied to Livestock Production*.
11. Henderson, C. R. (1973). SIRE EVALUATION AND GENETIC TRENDS. *Journal of Animal Science*, *1973*(Symposium), 10–41. https://doi.org/10.1093/ansci/1973.symposium.10
12. Himmelbauer, J., Schwarzenbacher, H., Fuerst, C., & Fuerst-Waltl, B. (2024). Exploring unknown parent groups and metafounders in single-step genomic best linear unbiased prediction: Insights from a simulated cattle population. *Journal of Dairy Science*, *107*(10), 8170–8192. https://doi.org/10.3168/jds.2024-24891
13. Kluska, S., Masuda, Y., Ferraz, J. B. S., Tsuruta, S., Eler, J. P., Baldi, F., & Lourenco, D. (2021). Metafounders may reduce bias in composite cattle genomic predictions. *Frontiers in Genetics*, *12*. https://doi.org/10.3389/fgene.2021.678587

14. Koivula, M., Strandén, I., Su, G., & Mäntysaari, E. (2012). Different methods to calculate genomic predictions—Comparisons of BLUP at the single nucleotide polymorphism level (SNP-BLUP), BLUP at the individual level (G-BLUP), and the one-step approach (H-BLUP). *Journal of Dairy Science*, *95*(7), 4065–4073. https://doi.org/10.3168/jds.2011-4874
15. Kudinov, A., Mäntysaari, E., Aamand, G., Uimari, P., & Strandén, I. (2020). Metafounder approach for single-step genomic evaluations of Red Dairy cattle. *Journal of Dairy Science*, *103*(7), 6299–6310. https://doi.org/10.3168/jds.2019-17483
16. Legarra, A., Bertrand, J., Strabel, T., Sapp, R., Sánchez, J., & Misztal, I. (2007). Multi-breed genetic evaluation in a Gelbvieh population. *Journal of Animal Breeding and Genetics*, *124*(5), 286–295. https://doi.org/10.1111/j.1439-0388.2007.00671.x
17. Legarra, A., Christensen, O. F., Aguilar, I., & Misztal, I. (2014). Single Step, a general approach for genomic selection. *Livestock Science*, *166*, 54–65. https://doi.org/10.1016/j.livsci.2014.04.029
18. Legarra, A., Christensen, O. F., Vitezica, Z. G., Aguilar, I., & Misztal, I. (2015). Ancestral relationships using metafounders: finite ancestral populations and across population relationships. *Genetics*, *200*(2), 455–468. https://doi.org/10.1534/genetics.115.177014
19. Liu, Z. (2011). Use of MACE results as input for genomic models. *Bulletin - International Bull Evaluation Service/Interbull Bulletin*, *43*. https://journal.interbull.org/index.php/ib/article/view/1176
20. Liu, Z., Goddard, M., Hayes, B., Reinhardt, F., & Reents, R. (2015). Technical note: Equivalent genomic models with a residual polygenic effect. *Journal of Dairy Science*, *99*(3), 2016–2025. https://doi.org/10.3168/jds.2015-10394

21. Liu, Z., Goddard, M., Reinhardt, F., & Reents, R. (2014). A single-step genomic model with direct estimation of marker effects. *Journal of Dairy Science*, *97*(9), 5833–5850. https://doi.org/10.3168/jds.2014-7924

22. Liu, Z., Reinhardt, F., Bünger, A., & Reents, R. (2004). Derivation and calculation of approximate reliabilities and Daughter Yield-Deviations of a random Regression Test-Day model for genetic evaluation of dairy cattle. *Journal of Dairy Science*, *87*(6), 1896–1907. https://doi.org/10.3168/jds.s0022-0302(04)73348-2

23. Lourenco, D., Legarra, A., Tsuruta, S., Masuda, Y., Aguilar, I., & Misztal, I. (2020). Single-Step Genomic Evaluations from Theory to Practice: Using SNP Chips and Sequence Data in BLUPF90. *Genes*, *11*(7), 790. https://doi.org/10.3390/genes11070790

24. Macedo, F., Astruc, J., Meuwissen, T., & Legarra, A. (2022). Removing data and using metafounders alleviates biases for all traits in Lacaune dairy sheep predictions. *Journal of Dairy Science*, *105*(3), 2439–2452. https://doi.org/10.3168/jds.2021-20860

25. Macedo, F. L., Christensen, O. F., Astruc, J., Aguilar, I., Masuda, Y., & Legarra, A. (2020). Bias and accuracy of dairy sheep evaluations using BLUP and SSGBLUP with metafounders and unknown parent groups. *Genetics Selection Evolution*, *52*(1), 47. https://doi.org/10.1186/s12711-020-00567-1

26. Mäntysaari, E. A., Evans, R. D., & Strandén, I. (2017). Efficient single-step genomic evaluation for a multibreed beef cattle population having many genotyped animals1. *Journal of Animal Science*, *95*(11), 4728–4737. https://doi.org/10.2527/jas2017.1912

27. Mäntysaari, E., Koivula, M., & Strandén, I. (2020). Symposium review: Single-step genomic evaluations in dairy cattle. *Journal of Dairy Science*, *103*(6), 5314–5326. https://doi.org/10.3168/jds.2019-17754

28. Mäntysaari, E., Liu, Z., & VanRaden, P. (2010). Interbull validation test for genomic evaluations. Bulletin - International Bull Evaluation Service. In *Interbull Bulletin, 41, 17*.

29. Masuda, Y. (2019). *Introduction to BLUPF90 suite programs.: Department of Animal and Dairy Science, University of Georgia*.

30. Masuda, Y., Misztal, I., Tsuruta, S., Legarra, A., Aguilar, I., Lourenco, D., Fragomeni, B., & Lawlor, T. (2016). Implementation of genomic recursions in single-step genomic best linear unbiased predictor for US Holsteins with a large number of genotyped animals. *Journal of Dairy Science*, *99*(3), 1968–1974. https://doi.org/10.3168/jds.2015-10540

31. Masuda, Y., Tsuruta, S., Bermann, M., Bradford, H. L., & Misztal, I. (2021). Comparison of models for missing pedigree in single-step genomic prediction. *Journal of Animal Science*, *99*(2). https://doi.org/10.1093/jas/skab019

32. Melo, T., Zwirtes, A., Silva, A., Lázaro, S., Oliveira, H., Silveira, K., Santos, J., Andrade, W., Kluska, S., Evangelho, L., Oliveira, H., & Tonhati, H. (2024). Unknown parent groups and truncated pedigree in single-step genomic evaluations of Murrah buffaloes. *Journal of Dairy Science*. https://doi.org/10.3168/jds.2023-24608

33. Misztal, I. (2015). Inexpensive Computation of the Inverse of the Genomic Relationship Matrix in Populations with Small Effective Population Size. *Genetics*, *202*(2), 401–409. https://doi.org/10.1534/genetics.115.182089

34. Misztal, I., Legarra, A., & Aguilar, I. (2014). Using recursion to compute the inverse of the genomic relationship matrix. *Journal of Dairy Science*, *97*(6), 3943–3952. https://doi.org/10.3168/jds.2013-7752

35. Misztal, I., Lourenco, D., & Legarra, A. (2020). Current status of genomic evaluation. *Journal of Animal Science*, *98*(4). https://doi.org/10.1093/jas/skaa101

36. Pocrnic, I., Lourenco, D. a. L., Bradford, H. L., Chen, C. Y., & Misztal, I. (2017). Technical note: Impact of pedigree depth on convergence of single-step genomic BLUP in a purebred swine population1. *Journal of Animal Science*, *95*(8), 3391–3395. https://doi.org/10.2527/jas.2017.1581

37. Ptak, E., Barć, A., & Jagusiak, W. (2015). *Rozwój metod oceny wartości hodowlanej zwierząt na przykładzie bydła mlecznego w ujęciu retrospektywnym*. Przegląd Hodowlany, 83(2).

38. Pyzara, A., Bylina, B., & Bylina, J. (2011). The influence of a matrix condition number on iterative methods' convergence. In *Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS): 18–21*.

39. R Core Team. (2023). *A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.* https://www.r-project.org/

40. Strandén, I. (2014). *RelaX2 program for pedigree Analysis, User's Guide for Version 1.65*.

41. Strandén, I., & Garrick, D. (2009). Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *Journal of Dairy Science*, *92*(6), 2971–2975. https://doi.org/10.3168/jds.2008-1929

42. Strandén, I., & Lidauer, M. (1999). Solving large mixed linear models using preconditioned conjugate gradient iteration. *Journal of Dairy Science*, *82*(12), 2779–2787. https://doi.org/10.3168/jds.s0022-0302(99)75535-9

43. Szyda, J., Żarnecki, A., Suchocki, T., & Kamiński, S. (2011). Fitting and validating the genomic evaluation model to Polish Holstein-Friesian cattle. *Journal of Applied Genetics*, *52*(3), 363–366. https://doi.org/10.1007/s13353-011-0047-z

44. Tsuruta, S., Misztal, I., Lourenco, D., & Lawlor, T. (2014). Assigning unknown parent groups to reduce bias in genomic evaluations of final score in US Holsteins. *Journal of Dairy Science*, *97*(9), 5814–5821. https://doi.org/10.3168/jds.2013-7821

45. Vandenplas, J., Calus, M. P. L., Eding, H., Van Pelt, M., Bergsma, R., & Vuik, C. (2021). Convergence behavior of single-step GBLUP and SNPBLUP for different termination criteria. *Genetics Selection Evolution*, *53*(1). https://doi.org/10.1186/s12711-021-00626-1

46. Vandenplas, J., Calus, M. P. L., Eding, H., & Vuik, C. (2019). A second-level diagonal preconditioner for single-step SNPBLUP. *Genetics Selection Evolution*, *51*(1). https://doi.org/10.1186/s12711-019-0472-8

47. Vandenplas, J., Eding, H., Calus, M. P. L., & Vuik, C. (2018). Deflated preconditioned conjugate gradient method for solving single-step BLUP models efficiently. *Genetics Selection Evolution*, *50*(1). https://doi.org/10.1186/s12711-018-0429-3

48. Vandenplas, J., Napel, J. T., Darbaghshahi, S. N., Evans, R., Calus, M. P. L., Veerkamp, R., Cromie, A., Mäntysaari, E. A., & Strandén, I. (2023). Efficient large-scale single-step evaluations and indirect genomic prediction of genotyped selection candidates. *Genetics Selection Evolution*, *55*(1), 37. https://doi.org/10.1186/s12711-023-00808-z

49. Vandenplas, J., Veerkamp, R., Calus, M., Lidauer, M., Strandén, I., Taskinen, M., Schrauf, M., & Napel, J. T. (2022). 358. MiXBLUP 3.0 – software for large genomic evaluations in animal breeding programs. *Proceedings of 12th World Congress on Genetics Applied to Livestock Production*, 1498–1501. https://doi.org/10.3920/978-90-8686-940-4_358

50. VanRaden, P. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*(11), 4414–4423. https://doi.org/10.3168/jds.2007-0980

51. Westell, R., Quaas, R., & Van Vleck, L. (1988). Genetic groups in an animal model. *Journal of Dairy Science*, *71*(5), 1310–1318. https://doi.org/10.3168/jds.s0022-0302(88)79688-5

8. Spis tabel i wykresów

Wykres 4. Średnia bezwzględna różnica między ostatecznym, a pośrednimi oszacowaniami wartości hodowlanych, oraz ich odchylenia standardowe, podczas procesu iteracyjnego dla zwierząt $G^{-}P^{-}$, $\boldsymbol{a}$ reprezentuje krowy, $\boldsymbol{b}$ reprezentuje buhaje.

Wykres 5. Średnia bezwzględna różnica w oszacowaniach efektów **SNP** między kolejnymi iteracjami, a ostatecznym wynikiem oraz ich odchylenia standardowe, dla różnego progu **MAF**.

Wykres 6. Kryterium zbieżności (CK, CM, CD) dla jednego i 12 rdzeni dla pełnego zbioru danych. Czarna linia przedstawia kryterium zatrzymania CD ≤ 1e-09.

Wykres 7. Kryteria zbieżności (CK, CM, CD) dla pełnego i zredukowanego zbioru danych. Czarna linia przedstawia kryterium zatrzymania CD ≤ 1e-09.

Wykres 8. Różnice pomiędzy średnimi wartościami oszacowań wartości hodowlanych pomiędzy podejściem **SNP-BLUP**, a podejściami z użyciem **G-BLUP**.

Wykres 9. Korelacja Pearsona pomiędzy pełnym, a zredukowanym zestawem danych dla poszczególnych scenariuszy.

Wykres 10. Ranking 50 buhajów o najwyższych wartościach hodowlanych.

Wykres 11. Ranking 50 najlepszych buhajów dla oszacowań wartości hodowlanych w obrębie każdego oprogramowania.

Wykres 12. Wyniki walidacji z podziałem na: płeć, status genotypowania, ilość brakujących danych rodowodowych oraz podejście kodowania brakujących informacji rodowodowych.

Wykres 13. Porównanie oszacowań wartości hodowlanych pomiędzy pełnym a zredukowanym zestawem danych dla krów z podziałem na status genotypowania, ilość brakujących danych rodowodowych oraz podejście kodowania brakujących informacji rodowodowych.

Wykres 14. Porównanie oszacowań wartości hodowlanych pomiędzy pełnym a zredukowanym zestawem danych dla buhajów z podziałem na status genotypowania, ilość brakujących danych rodowodowych oraz podejście kodowania brakujących informacji rodowodowych.

Wykres 15. Średni trend oszacowań wartości hodowlanych z podziałem na: płeć, ilość brakujących danych rodowodowych oraz podejście kodowania brakujących informacji rodowodowych.

9. Kopie publikacji wchodzących w skład rozprawy doktorskiej (załącznik nr 1)

10. Oświadczenia współautorów publikacji wchodzących w skład rozprawy doktorskiej (załącznik nr 2)

Genetics Selection Evolution

RESEARCH ARTICLE

Open Access

Heterogeneity in convergence behaviour of the single-step SNP-BLUP model across different effects and animal groups

Dawid Słomian[1], Kacper Żukowski[1] and Joanna Szyda[2*]

Abstract

Background The single-step model is becoming increasingly popular for national genetic evaluations of dairy cattle due to the benefits that it offers such as joint breeding value estimation for genotyped and ungenotyped animals. However, the complexity of the model due to a large number of correlated effects can lead to significant computational challenges, especially in terms of accuracy and efficiency of the preconditioned conjugate gradient method used for the estimation. The aim of this study was to investigate the effect of pedigree depth on the model's overall convergence rate as well as on the convergence of different components of the model, in the context of the single-step single nucleotide polymorphism best linear unbiased prediction (SNP-BLUP) model.

Results The results demonstrate that the dataset with a truncated pedigree converged twice as fast as the full dataset. Still, both datasets showed very high Pearson correlations between predicted breeding values. In addition, by comparing the top 50 bulls between the two datasets we found a high correlation between their rankings. We also analysed the specific convergence patterns underlying different animal groups and model effects, which revealed heterogeneity in convergence behaviour. Effects of SNPs converged the fastest while those of genetic groups converged the slowest, which reflects the difference in information content available in the dataset for those effects. Pre-selection criteria for the SNP set based on minor allele frequency had no impact on either the rate or pattern of their convergence. Among different groups of individuals, genotyped animals with phenotype data converged the fastest, while non-genotyped animals without own records required the largest number of iterations.

Conclusions We conclude that pedigree structure markedly impacts the convergence rate of the optimisation which is more efficient for the truncated than for the full dataset.

Background

The single-step model will soon become the standard procedure of most national genetic evaluations of dairy cattle [1, 2]. In spite of its great advantages for routine evaluations, with the most important being the possibility of conducting a joint breeding value estimation for ungenotyped and genotyped individuals, it should be kept in mind that statistically it is a very highly parameterised model that involves the estimation of several millions of effects that are often highly correlated. This poses potential problems in solving the system of equations, and most implementations have used the preconditioned

*Correspondence:
Joanna Szyda
joanna.szyda@upwr.edu.pl
[1] National Research Institute of Animal Production, Krakowska 1, 32-083 Balice, Poland
[2] Biostatistics Group, Department of Genetics, the Wroclaw University of Environmental and Life Sciences, Kozuchowska 7, 51-631 Wroclaw, Poland

conjugate gradient method (PCG) to solve these sets of equations. The PCG method was introduced for genetic evaluation models by Strandén and Lidauer in 1999 [3], which was further developed by Vandenplas et al. [4, 5] in the context of the single-step single nucleotide polymorphism best linear unbiased prediction (SNP-BLUP) model. Still, as indicated by these authors, especially in the context of the single-step SNP-BLUP model, a fast rate of convergence of the PCG requires an additional, second-level preconditioner. Moreover, on the national scale, the model is applied to very large datasets consisting of millions of records, which makes the solving technically and computationally demanding, not only in the context of memory consumption and CPU usage but also in the context of the numerical accuracy of calculations that is especially pronounced in parallel applications of PCG [6]. Relating to the above, the goal of our study was twofold: (1) to examine the differences in the model convergence depending on the number of individuals considered in the evaluation; and (2) to examine the convergence rate for different components of the model.

Methods

Materials

The analyzed dataset (Table 1) corresponds to the Polish national genetic evaluation for stature from December 2021 and comprised 1,098,611 cows with phenotypes for stature and 141,397 bulls with pseudo-phenotypes expressed by their de-regressed proofs (DRP) from the multiple across country evaluation (MACE) carried out by Interbull (interbull.org). Full genomic data in the form of genotypes of 46,118 SNPs were available for 134,960 individuals, including 70,134 cows with phenotypes as well as 64,826 bulls among which 26,471 were young individuals without pseudo-phenotypes and 38,355 were bulls with MACE-DRP. The majority of the genotyped individuals were genotyped using various versions of the EuroG MD Illumina genotyping microarray, containing more than 45,000 SNPs, that was customized for the EuroGenomics Cooperative. Individuals genotyped with other commercial platforms were imputed to EuroG MD using the Fimpute software [7]. In addition to the standard set of the 46,118 SNPs, for which a minor allele frequency (MAF) of 0.0064 corresponded to that of the genomic data used in the routine genomic evaluation, three SNP sub-sets, selected from this standard set, were considered based on MAF: (1) 45,537 SNPs with a MAF $\geq$ 0.01, (2) 41,667 SNPs with a MAF $\geq$ 0.05, and (3) 37,380 SNPs with a MAF $\geq$ 0.1. Furthermore, two strategies for the use of pedigree information were considered: (1) using the pedigree data for all available ancestors, i.e. 8,461,877 animals and 36 genetic groups (FULL); and (2) using the pedigree data for animals with phenotype or genotype data truncated after the fifth generation, which resulted in 1,555,995 individuals and 33 genetic groups (5GEN). Genotype data and pedigree information were stored in the cSNP database maintained by the National Research Institute of Animal Production [8].

Table 1 Numbers of animals in the analysed datasets

Category	Sex	Number of animals
Phenotype data	Females with phenotypes	1,098,611
	Males with MACE DRP	141,397
Genotype data	Females	70,134
	Males (bulls and candidates)	64,826
Pedigree data		
All generations	Females	6,428,481
	Males	2,023,328
5th generations	Females	1,368,487
	Males	187,508

Methods

The following single-step SNP-BLUP model [9] was considered:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Wa} + \mathbf{e}, \quad (1)$$

where $\mathbf{y}$ is the vector of dependent variables represented by the cows' measured phenotypes for stature and bulls' pseudo-phenotypes expressed by their MACE DRP, $\boldsymbol{\beta}$ is the vector of fixed effects including age at calving, lactation phase, and herd corresponding to the cows' phenotypes and bulls' DRP (note that for the bulls artificial codes for fixed effect class were used), $\mathbf{a}$ is the vector of the individuals' breeding values and genetic groups, which for the genotyped part of the population is expressed as $\mathbf{a} = \mathbf{Zg} + \mathbf{u}$ with $\mathbf{g}$ being the vector of random SNP effects, $\mathbf{u}$ is the vector of random additive (residual) polygenic effects, $\mathbf{e}$ is the vector of residuals, $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{W}$ are the corresponding design/incidence matrices. Genetic groups were defined based on a 10-year window based on the year of birth of the animals, separately for cows, bulls and their country of origin provided by: Poland, USA-Canada, and the remaining countries. The underlying covariance structure of the model is given by: $\mathbf{g} \sim \mathrm{MVN}\left(\mathbf{0}, \mathbf{I} \cdot \frac{1-k}{2\sum_{i=1}^{N} p_i(1-p_i)}\sigma^2_a\right)$, $\mathbf{u} \sim \mathrm{N}\left(\mathbf{0}, \mathbf{A} \cdot k\sigma^2_a\right)$, and $\mathbf{e} \sim \mathrm{N}\left(\mathbf{0}, \mathbf{D}\sigma^2_e\right)$, where $k(= 0.2)$ corresponds to the proportion of additive genetic variance due to the residual additive polygenic effect i.e. not explained by SNP genotype variation, p_i is the frequency of allele A of the ith SNP of N SNPs, $\mathbf{A}$ is the numerator relationship matrix constructed based on the pedigree information, and $\mathbf{D}$ is a diagonal matrix containing "1s" for cows with phenotypes or n_i for bulls with MACE DRP, where n_i

represents the MACE effective daughter contribution (EDC) [10]) for stature. $\sigma_a^2 = 5.50$ and $\sigma_e^2 = 4.63$ represent the additive polygenic and residual variance components, respectively. It should be noted that the variance components and the proportion of residual additive polygenic variance were not estimated, but were set to fixed values that corresponded to the parameters used in the Polish national genetic and genomic evaluation for stature.

Convergence

The effects of this model were estimated using the MiXBLUP software [11] that implements the two-level PCG [5] method for solving the following system of equations:

$$\mathbf{P}^{-1}\mathbf{M}^{-1}\mathbf{C}\mathbf{x} = \mathbf{P}^{-1}\mathbf{M}^{-1}\mathbf{b}, \quad (2)$$

where $\mathbf{C}$ is the coefficient matrix corresponding to the mixed model equations (MME) for solving Eq. (1), $\mathbf{x}$ is the vector of fixed and random effects given by $\mathbf{x}^{\mathrm{T}} = \left[\boldsymbol{\beta}^{\mathrm{T}}\mathbf{g}^{\mathrm{T}}\mathbf{u}^{\mathrm{T}}\right]$, $\mathbf{b}$ is the right hand side (RHS) of the MME, and $\mathbf{M}$ and $\mathbf{P}$ are the first level and the second level preconditioning matrices, respectively.

In our study, the convergence rate of $\mathbf{x}$ was expressed by three criteria: CK, CM, and CD, which is the relative absolute difference over the whole $\mathbf{x}$. The CK criterion was proposed by [12] as: $\frac{1}{\mu_1} \cdot \frac{\|\mathbf{M}^{-1}[\mathbf{b}-\mathbf{C}\hat{\mathbf{x}}_\mathbf{i}]\|}{\|\hat{\mathbf{x}}_\mathbf{i}\|}$, where μ_1 is the smallest active positive eigenvalue of the preconditioned coefficient matrix from Eq. (2) that influences the convergence and subscript i represents the round of iteration. The CM criterion was also proposed by [12] as: $\kappa(\mathbf{M}^{-1}\mathbf{C}) \cdot \frac{\|\mathbf{M}^{-1}[\mathbf{b}-\mathbf{C}\hat{\mathbf{x}}_\mathbf{i}]\|}{\|\mathbf{M}^{-1}\mathbf{b}\|}$, where $\kappa(\mathbf{M}^{-1}\mathbf{C})$ is the effective spectral condition number of the $\mathbf{M}^{-1}\mathbf{C}$ matrix. The CD criterion is given by: $\frac{\|\hat{\mathbf{x}}_{\mathbf{i}-1}-\hat{\mathbf{x}}_\mathbf{i}\|}{\|\hat{\mathbf{x}}_\mathbf{i}\|}$, where $\hat{\mathbf{x}}_\mathbf{i}$ represents estimates from the ith iteration. In our study, CD $\leq$ 1e − 09 was used as the stopping criterion, which indicates convergence of the equation system. The absolute difference is calculated as: $|\hat{\mathbf{x}}_\mathbf{i} - \hat{\mathbf{x}}_\mathbf{F}|$, where F corresponds to the estimates from the final iteration upon convergence. The absolute difference was calculated every 20th iteration, starting from the first iteration, separately for the following four groups of animals: (i) animals with genotypes and phenotypes $(\mathbf{G}^{+}\mathbf{P}^{+})$, including 59,242 animals, (ii) animals without genotypes and with phenotypes $(\mathbf{G}^{-}\mathbf{P}^{+})$, including 1,180,846 animals, (iii) animals with genotypes and without phenotypes $(\mathbf{G}^{+}\mathbf{P}^{-})$, including 75,718 animals, and (iv) animals without genotypes and phenotypes $(\mathbf{G}^{-}\mathbf{P}^{-})$, including 240,189 animals. Likewise, every 20th iteration, the absolute difference between estimated SNP effects was calculated.

The default stopping criterion ($\frac{\|\hat{\mathbf{x}}_{\mathbf{i}-1}-\hat{\mathbf{x}}_\mathbf{i}\|}{\|\hat{\mathbf{x}}_\mathbf{i}\|} \leq$ 1.0E–09), implemented into the MiXBLUP software, was used for the termination of the optimisation process. The optimisation and the corresponding results presented below were run in parallel using 12 cores. In addition, to assess the potential numerical instability of PCG due to parallel computations, the FULL model was also evaluated in serial execution using a single core.

Results

The convergence of the FULL dataset on 12 cores was achieved after 1240 iterations but when the model was run on a single core, it was reached 13 iterations later (Fig. 1). The 5GEN dataset with 682 iterations converged twice as fast (Fig. 2). Both convergence measures were very similar and demonstrated a nonlinear, even non-monotonical, improvement with an increase in CK and CM before reaching the final convergence. Pearson correlations between estimated breeding values (EBV) predicted based on those two datasets were very close to 1 regardless of the sex and animal group considered. As expected, the lowest correlations of 0.946 (bulls) and 0.980 (cows) were estimated for the least informative $(\mathbf{G}^{-}\mathbf{P}^{-})$ group (Table 2).

To explore the non-linearity of the convergence criteria in detail, the predicted breeding values (BV) and SNP effects obtained in each of the considered iterations were compared with their final values (Figs. 3 and 8). The three most striking features of the iterative process, regardless of the category considered, were (i) the convergence was reached twice as fast for the 5GEN than for the FULL dataset, (ii) as a result, the patterns of convergence expressed by the variability of some predicted values and by their differences to the final solution are much more pronounced in the 5GEN than in the FULL dataset, (iii) however, during iterations, before reaching convergence, the FULL dataset always resulted in estimates being considerably more similar to their final solutions than the 5GEN set, which may be due to differences in the preconditioning matrices $\mathbf{M}$ between both datasets. Regardless of the magnitude of the initial difference between EBV estimates in some iterations and the final estimate recorded for different groups, the convergence pattern is the same. After the initial phase showing a large variability in individual estimates, i.e. ~ the first 200 iterations, which sometimes even resulted in a relatively small averaged difference, the stable phase expressed by small differences in the accuracy of estimates across iterations was reached, followed by a rapid (5GEN) or monotonously decreasing (FULL) difference until the final estimate.

Considering the specific components of the random effects' solution vector, the EBV of animals with both

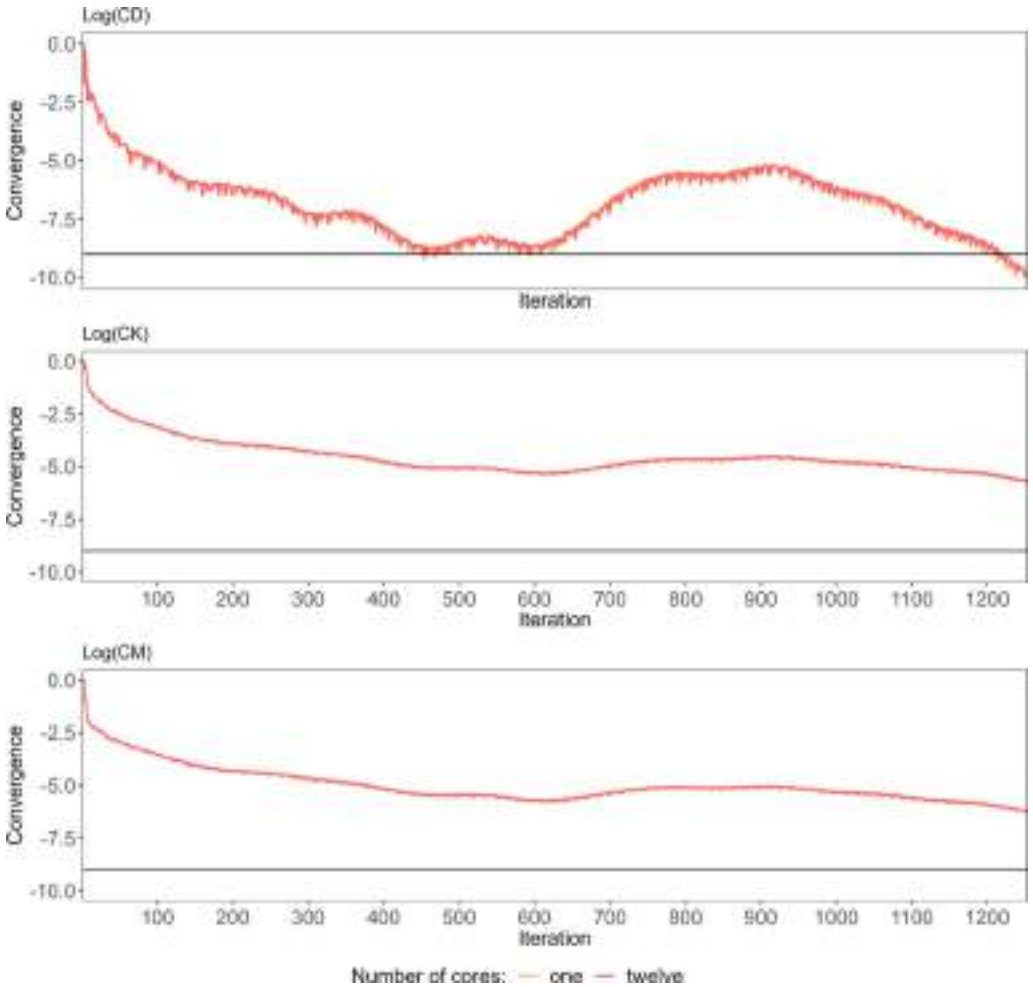


Fig. 1 The convergence criteria (CK, CM, CD) of PCG along the optimisation process. The convergence criteria for one and 12 cores for the FULL dataset. The black line represents the stopping criterion defined by CD ≤ 1e−09

sources of information available $(\mathbf{G}^{+}\mathbf{P}^{+})$ demonstrated the smallest absolute differences between the final EBV and the EBV estimated during the optimisation iterations; in addition, their standard deviations (especially for the bulls) were the smallest of all considered groups of animals, which may indicate good quality starting values for the iteration process (Fig. 3) and higher reliabilities of the EBV of the animals as compared to the other groups. For the $(\mathbf{G}^{-}\mathbf{P}^{+})$ group, the pattern and average absolute difference in EBV convergence were very similar to those of the $(\mathbf{G}^{+}\mathbf{P}^{+})$ group, but a much larger variability of some predicted EBV was observed for bulls (Fig. 4). The same results were found for the $(\mathbf{G}^{+}\mathbf{P}^{-})$ group (Fig. 5). In spite of the similar convergence pattern as expressed by the average absolute difference, during almost the full process of iterations a considerably larger variability in the predicted EBV was observed for some individuals from the least informative $(\mathbf{G}^{-}\mathbf{P}^{-})$ group. For this group, the EBV reliabilities were the lowest among the four groups, which indicates that certain members of this group are the limiting factor that affects convergence (Fig. 6). Unlike the EBV, the estimation of the effects of genetic groups revealed a monotonically improving pattern of convergence. However, the most striking feature was the heterogeneity in convergence behaviour between the two datasets and some groups. Although the estimates of some genetic groups of the 5GEN dataset already reached convergence after the 200th iteration, for the FULL dataset the convergence of the estimates was poor (Fig. 7). The opposite convergence behaviour was attributed to SNP effects that, regardless of the dataset, converged very fast, so that final estimates were already available at the 300th iteration, i.e. in the middle (5GEN) or even at one-third (FULL) of the whole optimisation process

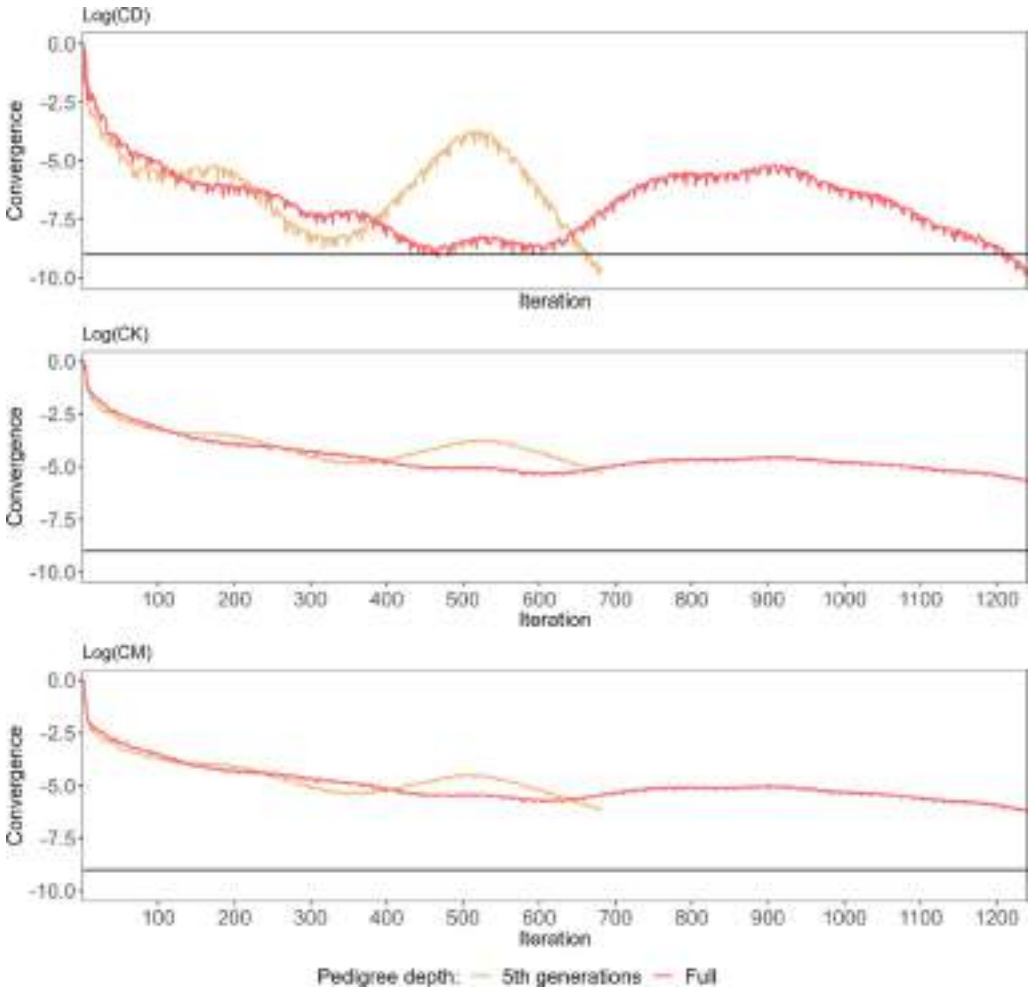


Fig. 2 The convergence criteria (CK, CM, CD) of PCG along the optimisation process. The convergence criteria for the FULL and 5GEN datasets. The black line represents the stopping criterion defined by CD ≤ 1e−09

Table 2 Correlations between EBV predicted based on the full pedigree and on the pedigree truncated after the 5th generation for different groups of animals

Group of animals	Number of animals	Correlation	
		Males	Females
All	1,555,995	0.997	0.991
Phenotype and genotype	59,242	0.999	0.999
Only phenotype	1,180,846	0.999	0.999
Only genotype	75,718	0.999	0.999
Without genotype and phenotype	240,189	0.946	0.980

(Fig. 8). The correlation between the final estimates upon convergence of the FULL dataset resulting from parallel and linear computations was 0.99, with some differences occurring from the third decimal place. Conversely, the four SNP subsets demonstrated very similar convergence rates and patterns (Fig. 9).

Finally, we compared the 50 top EBV ranking bulls between the 5GEN and FULL datasets and found 49 overlapping bulls. The single bull from the 5GEN data that was not in the top 50 bulls of the FULL dataset was still classified at the relatively high 52nd rank. In contrast, one bull that ranked high in the FULL dataset was missing in the top 50 bulls in 5GEN dataset, this bull was not evaluated in Poland, and its high predicted EBV in the FULL dataset was due to several highly ranked relatives that were present in the full pedigree records. The overall rank correlation for the remaining 49 individuals was 0.99.

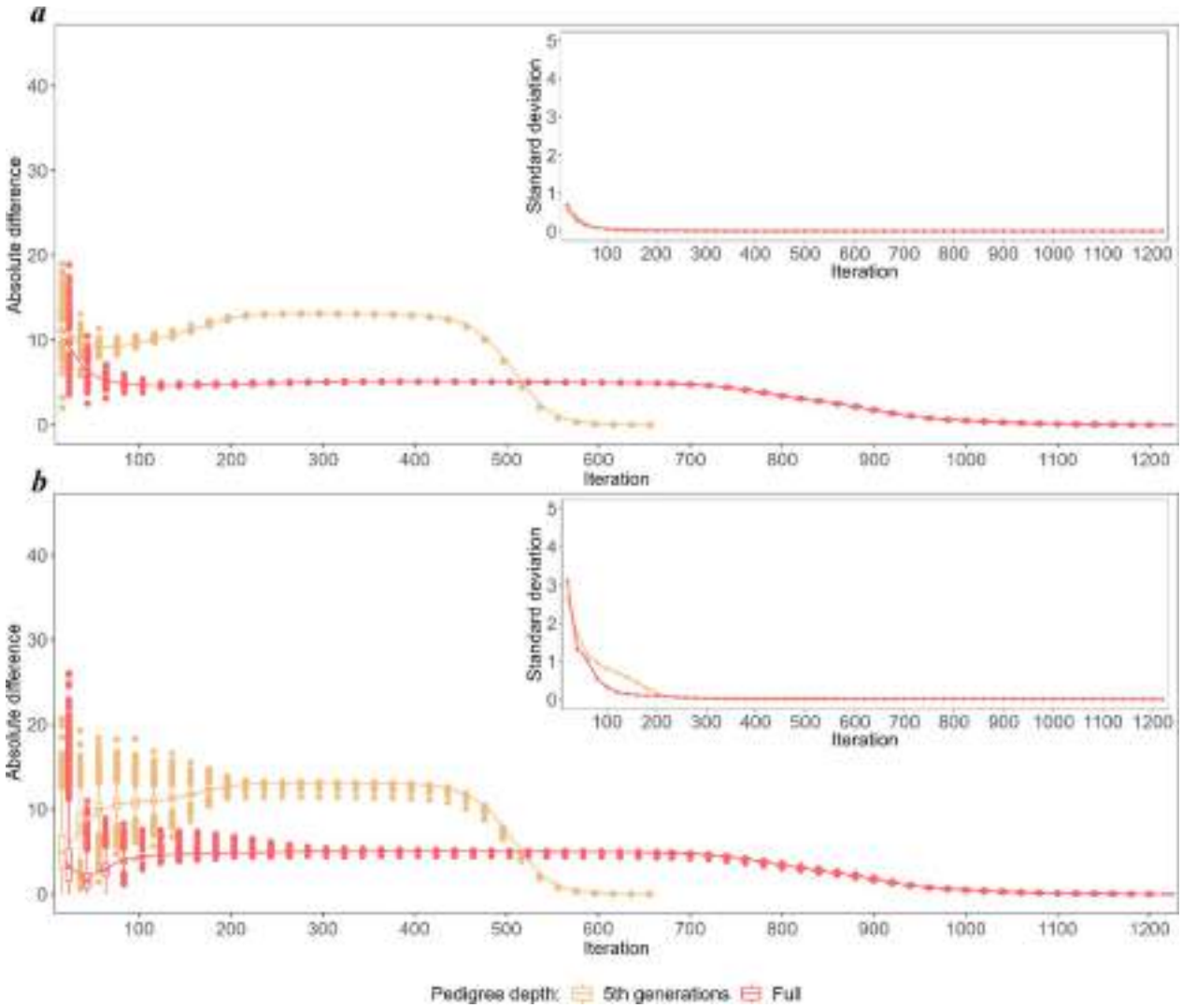


Fig. 3 The average absolute difference in estimated breeding values (EBV) between a given round of iteration and the final solution (main graph) and their standard deviations (inside graph) during the optimisation process for animals with phenotypes and genotypes ($G^{+}P^{+}$). **a** Shows the convergence criteria for the FULL and 5GEN datasets. **b** Shows the convergence criteria for one and 12 cores for the FULL dataset

Discussion

The most striking result of our study was the much faster convergence of the reduced (5GEN) dataset than that of the full dataset, although solutions from the initial rounds of iterations were much better (i.e. closer to the final solutions) for the FULL dataset. Although in our study, this result was obtained by solving a single-step SNP-BLUP model, Pocrnic et al. [13] reported the same result with the single-step genomic BLUP (GBLUP) model. In addition, Legarra et al. [1] observed potential convergence problems for data structures composed of a deep pedigree with many generations of ungenotyped individuals. Another common feature of the solving process was a non-linear and even non-monotonic convergence pattern. Although Vandenplas et al. [12] indicated an approximately linear convergence expressed by the CK, CM, and CD criteria, our study as well as other applications of PCG to the solving of mixed linear models in the context of SNP-BLUP and GBLUP [4, 13–16] reported a pattern that approximates a typical nonlinear behaviour that involves an initial phase of fast convergence rate, a second phase characterised by an almost linear convergence rate, and a third fast converging phase, although with a non-monotonic decrease of the convergence measures. However, for most of the considered scenarios, the initial phase did not always show a rapid linear convergence of estimates, as expressed by their absolute difference from the final solution, and

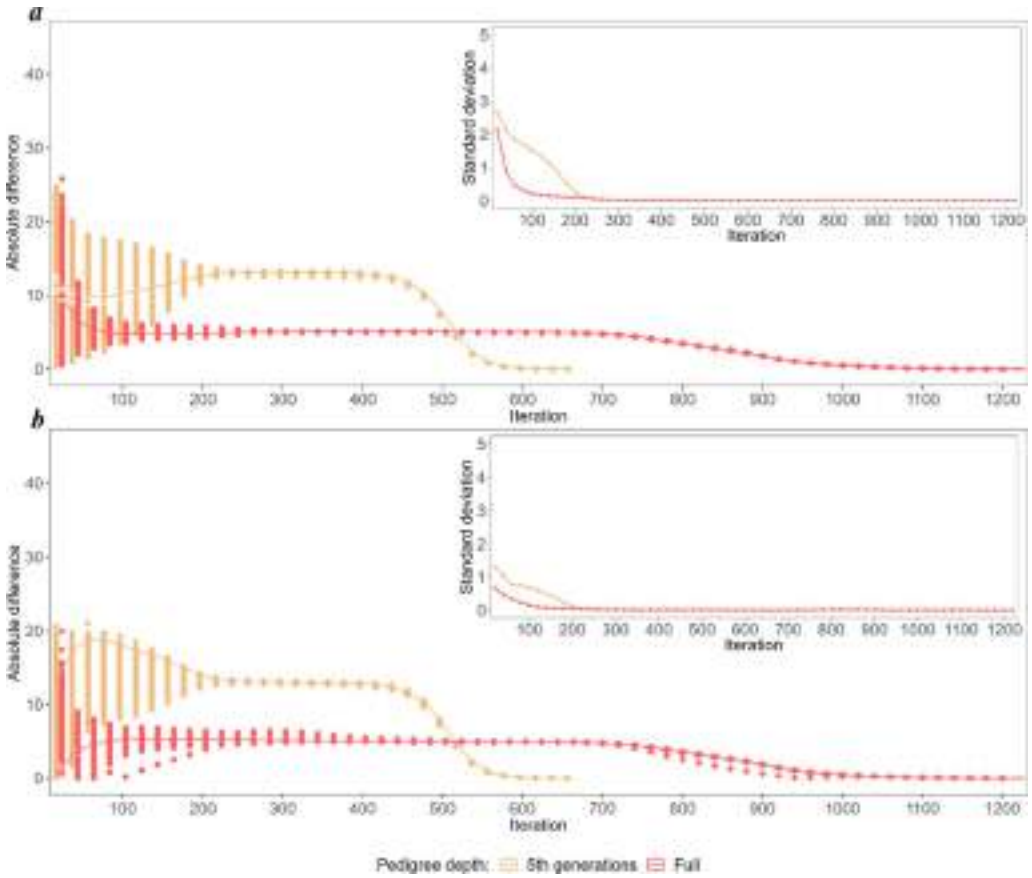


Fig. 4 The average absolute difference in estimated breeding values (EBV) between a given round of iteration and the final solution (main graph) and their standard deviations (inside graph) during the optimisation process for ungenotyped animals with phenotypes $(G^{-}P^{+})$. **a** Represents bulls and **b** Represents cows

involved a small number of iterations. Furthermore, we observed differences in the initial convergence rate that varied not only between the two datasets (FULL vs. 5GEN) but also between the groups of effects considered (EBV for $\mathbf{G}^{+}\mathbf{P}^{+}$, $\mathbf{G}^{+}\mathbf{P}^{-}$, $\mathbf{G}^{-}\mathbf{P}^{+}$, $\mathbf{G}^{-}\mathbf{P}^{-}$; genetic groups, and SNPs). We hypothesise that the differences between datasets are due to the pedigree structure, which results in a smaller number of individuals without genotype and phenotype information. Furthermore, the differences between animal groups are due to different information contents that are implemented into the preconditioner matrix $\mathbf{M}$ and that impact the quality of preconditioning, as defined in [5], or to the effects of genetic groups and breeding values of the $\mathbf{G}^{-}\mathbf{P}^{-}$ individuals, which are predicted only indirectly based on relatives with genotypic and/or phenotypic data. An interesting convergence pattern was observed when comparing the FULL and the 5GEN datasets, i.e. the second (linear) convergence phase was always much longer with the FULL dataset, which may result from the fact that equation systems of large dimensions are typically not as well conditioned as smaller equation systems, which impacts the efficiency of iterative solvers [17]. Indeed, in the case of our data, the condition number, computed as the ratio of the approximated extremal eigenvalues, was twice as high for the FULL dataset (4,258 285) than for the 5GENE dataset (2,171,642), resulting in a smaller effective spectral condition number that can be related with a faster convergence. In addition, numerical instability, which is due to rounding errors since the evaluation of a complete pedigree involves many more arithmetic operations, may further hamper the numerical performance [4, 6].

In the original application of the PCG algorithm for the optimisation of the single-step BLUP evaluation, it was observed that SNP effects posed a problem for the efficient convergence of the solver [4]. However, in the current evaluation, among all the effects, SNP estimates demonstrated the fastest convergence from the beginning of the iteration and then, by far, the fastest and nearly linear convergence towards the final solutions. On the one hand, this is due to the implementation of an additional pre-conditioning matrix (the second-level

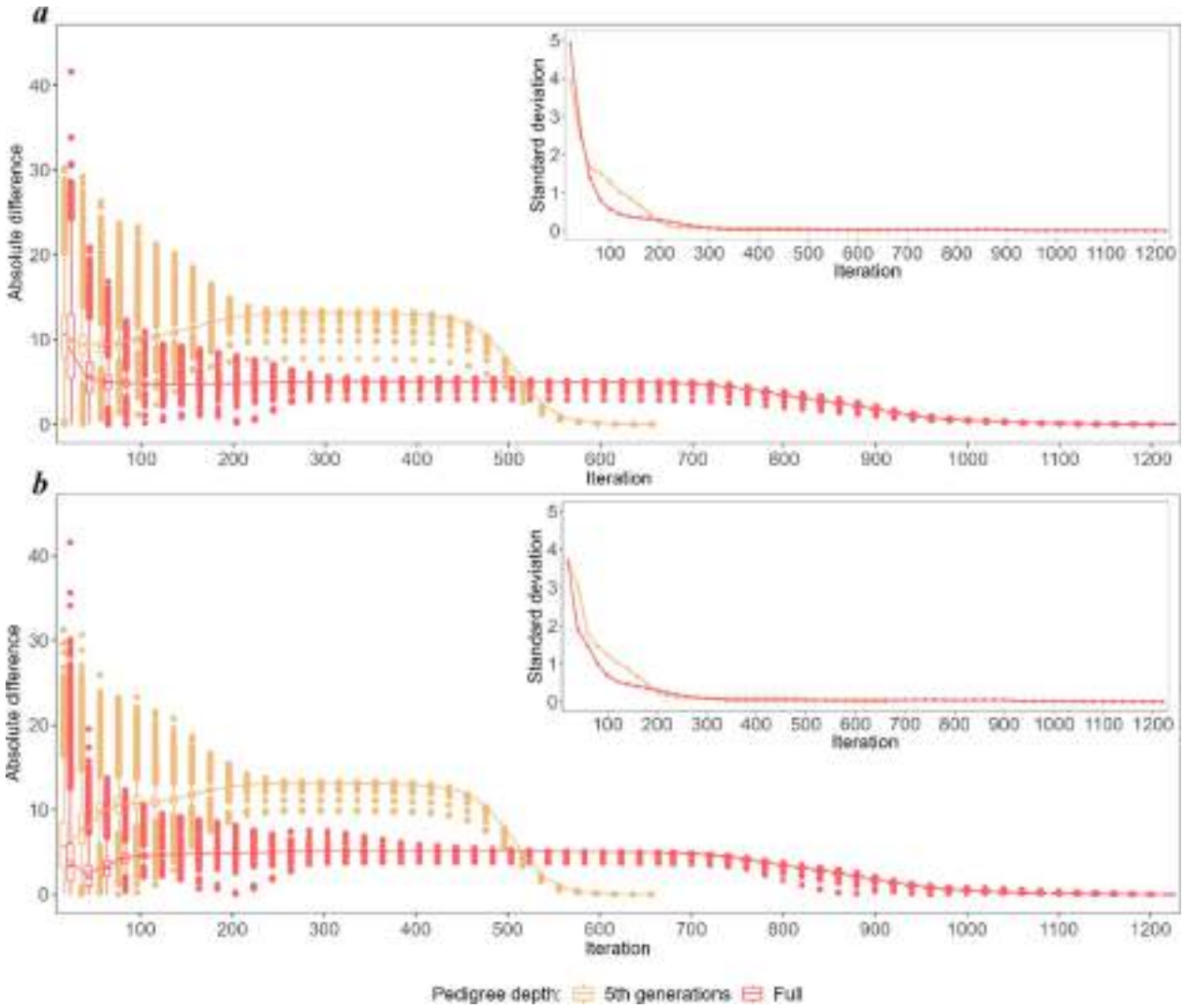


Fig. 5 The average absolute difference in estimated breeding values (EBV) boxplots between a given round of iteration and the final solution (main graph) and their standard deviations (inside graph) during the optimisation process for genotyped animals without phenotypes $\left(G^{+}P^{-}\right)$. **a** Represents bulls, and **b** Represents cows

preconditioner **P**) into the current version of the MiXBLUP software. On the other hand, a smaller number of genotyped individuals (compared to the number of phenotyped individuals) available for the estimation of SNP effects corresponds to a smaller number of arithmetic operations involved, which implies a smaller number of rounding errors by storing and processing real type variables involved in the computations [4]. The third observation was that the faster convergence of SNP effects may also be due to more information content being available for these in the dataset, since the EBV of all genotyped animals make contributions to the estimation of SNP effects.

Finally, we analysed potential convergence problems that can originate from the additional numerical complexity imposed due to the implementation of parallel computations, as indicated by [6]. However, this phenomenon was not observed in the single-step SNPBLUP implementation.

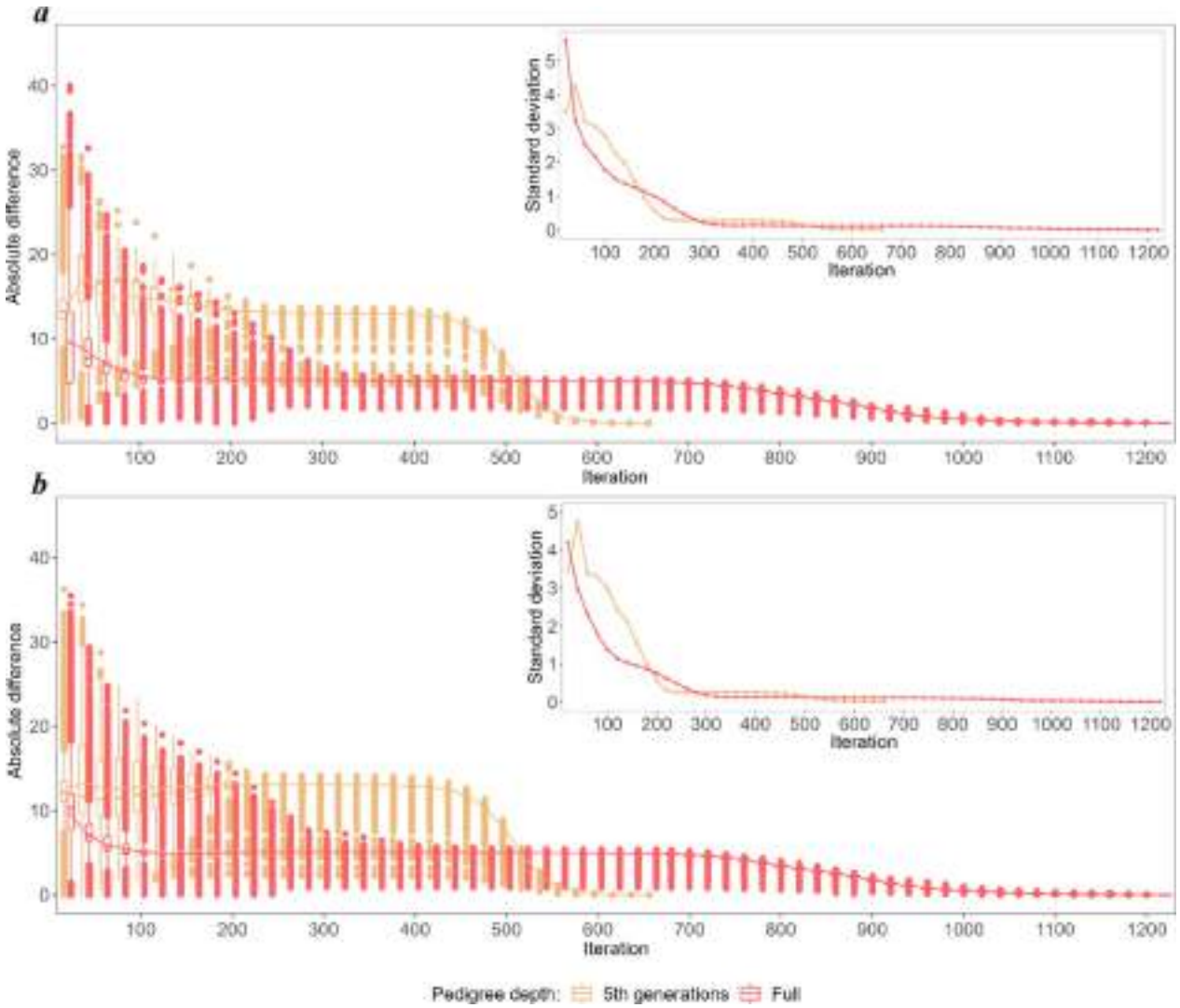


Fig. 6 The average absolute difference in estimated breeding values (EBV) between a given round of iteration and the final solution (main graph) and their standard deviations (inside graph) during the optimisation process for ungenotyped animals without phenotypes ($G^{-}P^{-}$). **a** Represents bulls, and **b** Represents cows

Conclusions

Our findings demonstrate that not all effects estimated in the single-step SNP-BLUP model have the same convergence rate. The effects of SNPs converged the fastest, whereas those of the genetic groups had the lowest rate of convergence. Among the four groups of animals, the EBV of the genotyped animals with phenotype data reached the final solutions with the smallest number of iterations, whereas the nongenotyped animals without own phenotype records required the largest number of rounds of iterations to reach their final solutions. The depth of pedigree markedly influences the rate of convergence, with fewer generations in the pedigree leading to faster convergence. We believe, that the observed convergence patterns are not specific to the dataset analyzed here, which reflects a structure of a standard national dairy population, but that they also apply to other national populations.

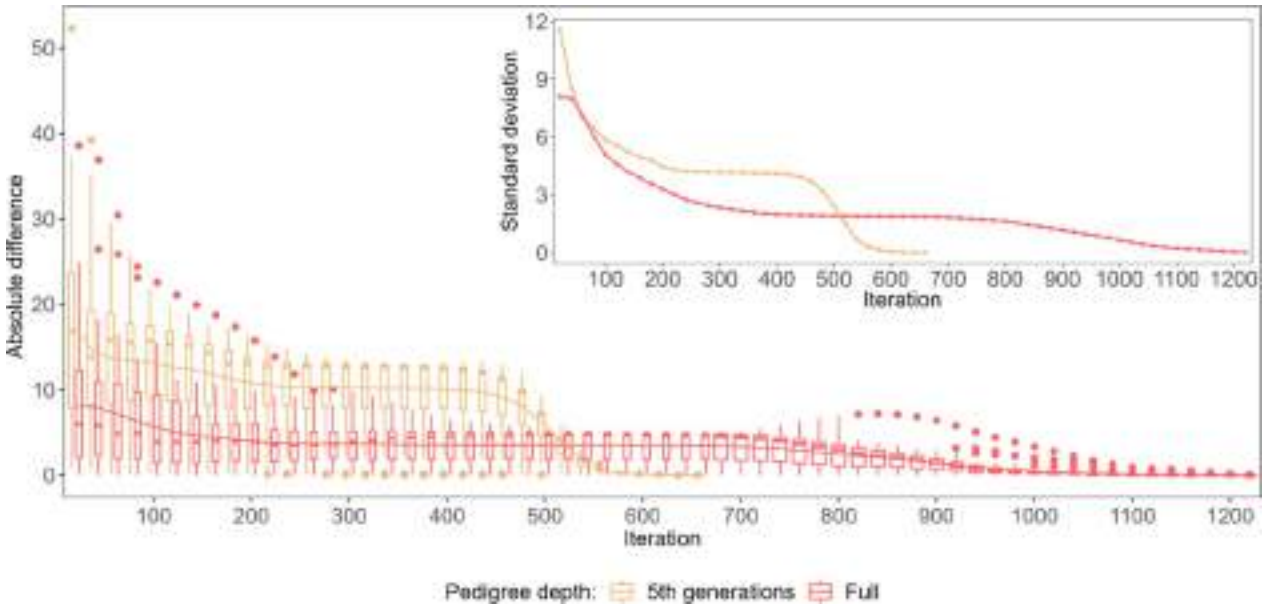


Fig. 7 The average absolute difference in the estimates of genetic groups effects between a given round of iteration and the final solution (main graph) and their standard deviations (inside graph) during the optimisation process

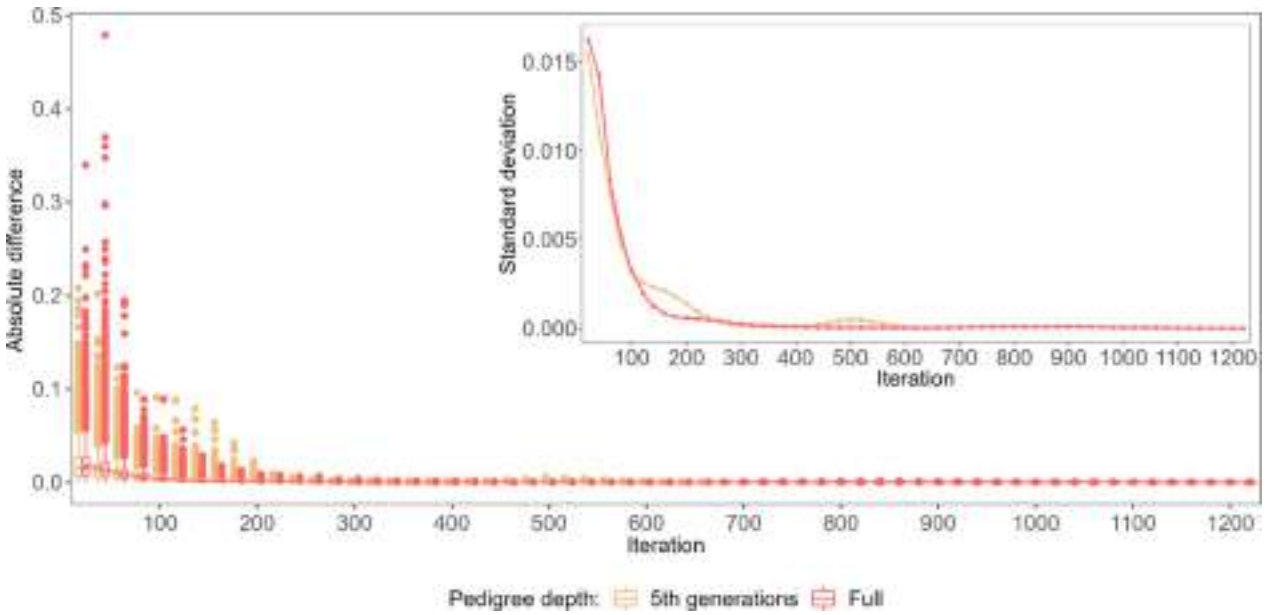


Fig. 8 The average absolute difference in the estimates of SNP effects between the particular iteration and the final solution (main graph) and their standard deviations (inside graph) during the optimisation process

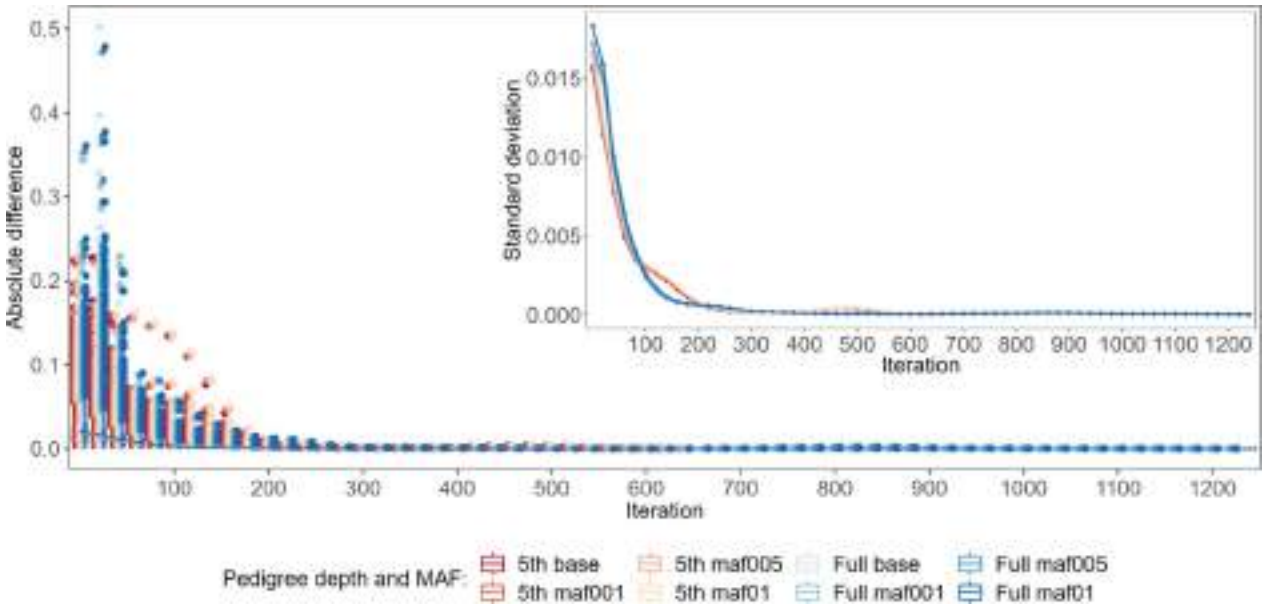


Fig. 9 The average absolute difference in the estimates of SNP effects between the particular iteration and the final solution (main graph) and their standard deviations (inside graph) during the optimisation process for different SNP preselection criteria based on the MAF threshold

Acknowledgements
The authors thank anonymous reviewers since their comments significantly contributed to the scientific quality of the text.

Authors' contributions
DS performed all data analyses, and was involved in creating the study concept, KŻ prepared and edited genomic data, JS conceptualised the study, interpreted the data and drafted the manuscript. All authors contributed to the writing of the manuscript and reviewed the final version. All the authors read and approved the final manuscript.

Funding
This study was supported by a grant from the Ministry of Agriculture and Rural Development (DŻW.pp.862.1.2023).

Availability of data and materials
The data set corresponding to the genetic evaluation from December 2021 can be obtained from the Institute of Animal Breeding on request to ocena.bydlo@iz.edu.pl.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 13 April 2023 Accepted: 15 November 2023
Published online: 23 November 2023

References

1. Legarra A, Christensen OF, Aguilar I, Misztal I. Single step, a general approach for genomic selection. Livest Sci. 2014;166:54–65.
2. Mäntysaari EA, Evans RD, Strandén I. Efficient single-step genomic evaluation for a multibreed beef cattle population having many genotyped animals. J Anim Sci. 2017;95:4728–37.
3. Strandén I, Lidauer M. Solving large mixed linear models using preconditioned conjugate gradient iteration. J Dairy Sci. 1999;82:2779–87.
4. Vandenplas J, Eding H, Calus MPL, Vuik C. Deflated preconditioned conjugate gradient method for solving single-step BLUP models efficiently. Genet Sel Evol. 2018;50:51.
5. Vandenplas J, Calus MPL, Eding H, Vuik C. A second-level diagonal preconditioner for single-step SNPBLUP. Genet Sel Evol. 2019;51:30.
6. Cools S, Fatih Yetkin E, Agullo E, Giraud L, Vanroose W. Analyzing the effect of local rounding error propagation on the maximal attainable accuracy of the pipelined conjugate gradient method. SIAM J Matrix Anal Appl. 2018;39:426–50.
7. Sargolzaei M, Chesnais JP, Schenkel FS. A new approach for efficient genotype imputation using information from relatives. BMC Genomics. 2014;15:478.
8. Żukowski K, Makarski J, Mazanek K, Prokowski A. Database management of single nucleotide polymorphism for use in Polish genomic evaluation. In: Proceedings of the 66th annual meeting of the European federation of animal science: 31 August–4 September 2015;2015, Warsaw.
9. Liu Z, Goddard ME, Reinhardt F, Reents R. A single-step genomic model with direct estimation of marker effects. J Dairy Sci. 2014;97:5833–50.
10. Liu Z. Use of MACE results as input for genomic models. Interbull Bull. 2011;43:1–4.
11. Ten Napel J, Vandenplas J, Lidauer M, Stranden I, Taskinen M, Mäntysaari E, et al. MiXBLUP 2.2.0 manual. 2020. https://www.mixblup.eu/documents/RvT_06335_ASG_WLR_MixBlup%20Manual_LR-spread.pdf/. Accessed 31 Oct 2023.
12. Vandenplas J, Calus MPL, Eding H, van Pelt M, Bergsma R, Vuik C. Convergence behavior of single-step GBLUP and SNPBLUP for different termination criteria. Genet Sel Evol. 2021;53:34.
13. Pocrnic I, Lourenco DAL, Bradford HL, Chen CY, Misztal I. Technical note: Impact of pedigree depth on convergence of single-step genomic BLUP in a purebred swine population. J Anim Sci. 2017;95:3391–5.

14. Vandenplas J, Eding H, Calus MPL. Technical note: Genetic groups in single-step single nucleotide polymorphism best linear unbiased predictor. J Dairy Sci. 2021;104:3298–303.
15. Strandén I, Aamand GP, Mäntysaari EA. Single-step genomic BLUP with genetic groups and automatic adjustment for allele coding. Genet Sel Evol. 2022;54:38.
16. Harris BL, Sherlock RG, Nilforooshan MA. Large-scale multiple-trait single-step marker model implementation. In: Proceedings of 12th World Congress on Genetics Applied to Livestock Production: 3–8 July 2022; Rotterdam. 2022.
17. Pyzara A, Bylina B, Bylina J. The influence of a matrix condition number on iterative methods' convergence. In: Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS): 18–21 September 2011; Szczecin. 2011.

Publisher's Note

ANNALS OF ANIMAL SCIENCE
ISSN: 2300-8733, https://sciendo.com/journal/AOAS

ACCEPTED AUTHOR VERSION OF THE MANUSCRIPT:

A comparison of genomically enhanced breeding values predicted by different single-step approaches

DOI: 10.2478/aoas-2025-0088

Dawid Słomian[1], Kacper Żukowski[1], Joanna Szyda[2,1]♦

[1]Department of Cattle Breeding, National Research Institute of Animal Production, 32-083 Balice n. Kraków, Poland

[2]Department of Genetics, The Biostatistic Group, Wroclaw University of Environmental and Life Sciences, Kozuchowska 7, 51-631 Wrocław, Poland

♦Corresponding author: joanna.szyda@upwr.edu.pl

Received date: 15 February 2025
Accepted date: 8 August 2025

To cite this article: (2025). Słomian D., Żukowski K., Szyda J. (2025). A comparison of genomically enhanced breeding values predicted by different single-step approaches, Annals of Animal Science, DOI: 10.2478/aoas-2025-0088

A comparison of genomically enhanced breeding values predicted by different single-step approaches

Dawid Słomian[1], Kacper Żukowski[1], Joanna Szyda[2,1♦]

[1]Department of Cattle Breeding, National Research Institute of Animal Production, 32-083 Balice n. Kraków, Poland

[2]Department of Genetics, The Biostatistic Group, Wroclaw University of Environmental and Life Sciences, Kozuchowska 7, 51-631 Wrocław, Poland

♦Corresponding author: joanna.szyda@upwr.edu.pl

DOI: 10.2478/aoas-2025-0088

Abstract

Many countries are currently adopting the single-step model for national genetic evaluations of dairy cattle. The two most widely applied statistical formulations of the single-step model are Genomic Best Linear Unbiased Prediction (ssG-BLUP) and Single Nucleotide Polymorphism BLUP (ssSNP-BLUP), with the main difference being the handling of additive genetic covariance between individuals with genotypes. Using solvers available in MiXBLUP software, our study aimed to compare both models in terms of Genomic Enhanced Breeding Value (GEBV) prediction, bull rankings, and computational efficiency (memory consumption and computational time). The results did not show marked differences in the quality of GEBV prediction expressed by the metrics underlying the Interbull validation, except for ssG-BLUP, APY-based solvers with 3,000 core bulls. However, the ranking of the top 50 bulls differed between models, which has implications for the breeding industry and selection, since the top-ranking bulls are typically the most widely used. 39 and 31 of the top 50 bulls were common to all models for stature and foot angle, respectively. Regarding computational time, ssSNP-BLUP and ssG-BLUP with APY solver using 3,000 bulls were the fastest, and ssG-BLUP with GT solver was the slowest. The selection of core individuals for the APY solver was a crucial element that affected the prediction accuracy. GT or SNP-BLUP solvers can circumvent this issue, since no selection of core individuals is required.

Key words: APY, G-BLUP, GEBV, GT, single-step, SNP-BLUP

Abbreviations

APY – algorithm for proven and young

BLUP – best linear unbiased prediction

DGV – direct genetic value

DRP – deregressed proofs

EDC – effective daughter contributions

GEBV – genomically enhanced breeding values

h^2 – heritability

MACE – multiple across country evaluation

MME – mixed model equations

RHS – right hand side

SNP – single nucleotide polymorphisms

Today, many countries are implementing the single-step model for their national genetic evaluations of dairy cattle. The main difference between the currently used multi-step and single-step models is the incorporation of genomic information available for calculating additive genetic covariances between individuals using phenotype, genotype, and pedigree information simultaneously. The single-step model incorporates the identical-by-descent or at least identical-by-state similarity between the genotyped fraction of the evaluated individuals, expressed by the similarity of their genotypes. Genotypic information is typically expressed by single nucleotide polymorphisms (SNPs). Two equivalent statistical formulations of such a single-step model comprise the genomic best linear unbiased prediction (ssG-BLUP) [1, 2, 3] and ssSNP-BLUP. The only difference between these models lies in the formulation of the additive genetic covariance between individuals, in particular, between individuals with genotypes. In ssG-BLUP, the covariance matrix of breeding values of all evaluated individuals comprises the components related to non-genotyped individuals constructed using the pedigree information, and to genotyped animals constructed as a weighted sum of pedigree and SNP genotype information. In the ssSNP-BLUP, the genomic relationship between the additive genetic effects of SNPs is incorporated as a separate component in addition to a pedigree-based relationship. However, despite different parameterizations both models are equivalent mathematically [4, 5].

However, practical applications for the evaluation of large national populations of dairy cattle are computationally very challenging. Three solvers have been developed and are used for national genetic evaluations. The ssG-BLUP model can be implemented either by: the APY solver [6] or the GT solver [6]. The APY (algorithm for proven and young) procedure does not use the entire genomic relationship matrix but only a part of it based on a predefined set of genotyped individuals (so-called core individuals) while the remainder, i.e. non-core genotyped individuals is fitted as genomically uncorrelated using a diagonal genomic relationship matrix. Therefore, the system for solving model equations only requires a computationally intensive inverse of the dense genomic submatrix for the core individuals and the submatrices corresponding to the covariance between core and noncore individuals. The GT procedure implements an inverse of the genomic relationship matrix as a function of the inverse of the pedigree-based numerator relationship matrix corresponding to genotyped individuals and the

SNP genotype design matrix, calculated based on the Woodbury formula. In the ssSNP-BLUP [4], due to an alternative formulation of the covariance structure, solving the system of equations does not require inverting of the partial (APY) or full (GT) relationship matrices but only a computationally simple inverse of the SNP covariance matrix.

The purpose of this study was to compare G-BLUP, GT, and SNP-BLUP solvers in terms of model validation performance to predict breeding values, differences in predicted breeding values between solvers, as well as their computational efficiency and computational resources requirements, in the context of routine genetic evaluation.

Material and methods

Material

The analyzed set (Table 1) represented data from the Polish national genetic evaluation of Holstein cattle from December 2021 for stature, which represents a highly heritable trait (h^2=0.54), and foot angle, representing a low heritability trait (h^2=0.09). It comprised 1,098,611 cows with stature phenotypes and 1,098,766 cows with foot angle phenotypes born between 1992 and 2019 as well as 141,397 bulls (stature) and 117,482 bulls (foot angle) born between 1986 and 2017 with pseudophenotypes expressed by their deregressed proofs (DRP) from the multiple across country evaluation (MACE) carried out by Interbull (interbull.org). Genomic data in the form of 46,118 SNP markers were available for 134,960 individuals, including 70,134 cows with phenotypes born between 2009 and 2021, as well as 64,826 bulls born between 1985 and 2021, of which 26,471 represented young individuals without pseudophenotypes and 38,355 were bulls with MACE-DRP. In our study, the pedigree of animals with phenotype data was truncated after the fifth generation, resulting in 1,555,995 individuals and 33 genetic groups.

Table 1. Number of animals in the analyzed data sets

Category	Sex	Number of animals
Phenotype data (Stature)	Females with phenotypes	1,098,611
	Males with MACE DRP	141,397
Phenotype data (Foot angle)	Females with phenotypes	1,098,766
	Males with MACE DRP	117,482
Genotype data	Females	70,134
	Males (bulls and candidates)	64,826
Pedigree data	Females	1,368,487
	Males	187,508

GEBV prediction models

The following single-step single-trait models were considered for the prediction of GEBV.

The ssG-BLUP [2]:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{W}_G\boldsymbol{a} + \boldsymbol{e}, \quad (1)$$

where $\boldsymbol{y}$ is the vector of dependent variables represented by cows' measured phenotypes and bulls' pseudo phenotypes expressed by their MACE DRPs, **b** represents a vector of fixed effects scored at cows' first parity including age at calving, lactation phase, and herd, while for bulls the corresponding fixed effects were represented by assigning common classes, $\boldsymbol{a}$ represents a vector of breeding values, and $\boldsymbol{e}$ is the vector of residuals. The underlying covariance structure of the random effects is given by $\boldsymbol{a} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{H}_G\sigma_a^2)$ and $\boldsymbol{e} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{R}\sigma_e^2)$. $\boldsymbol{H}_G$ is given by $\begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} + \begin{bmatrix} \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}(\boldsymbol{G} - \boldsymbol{A}_{22})\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} & \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}(\boldsymbol{G} - \boldsymbol{A}_{22}) \\ (\boldsymbol{G} - \boldsymbol{A}_{22})\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} & \boldsymbol{G} - \boldsymbol{A}_{22} \end{bmatrix}$ [7], where $\boldsymbol{A}_{11}$, $\boldsymbol{A}_{12}/\boldsymbol{A}_{21}$, and $\boldsymbol{A}_{22}$ are the components of the numerator relationship matrix constructed based on the pedigree corresponding to non-genotyped animals, the covariance between non-genotyped and genotyped animals, as well as between genotyped animals, respectively, while $\boldsymbol{G}$ represents the genomic relationship matrix between genotyped animals. $\boldsymbol{R}$ is a diagonal matrix containing 1.00 for cows with phenotypes or n_i for bulls with MACE DRPs, with n_i representing a difference in effective daughter contributions of i-th bull between the MACE and the national evaluation. $\boldsymbol{X}$ and $\boldsymbol{W}_G$ denote the corresponding design matrices.

The ssSNP-BLUP [4]:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{W}_S\boldsymbol{a} + \boldsymbol{e}, \quad (2)$$

where $\boldsymbol{y}$, $\boldsymbol{b}$, $\boldsymbol{a}$, and $\boldsymbol{e}$ are the same as defined above, but $\boldsymbol{a}$ is parameterized as $\boldsymbol{Zg} + \boldsymbol{u}$ with $\boldsymbol{g}$ being the vector of random SNP effects and $\boldsymbol{u}$ being- a vector of random additive residual polygenic effects. The covariance structure imposed on the residual effect ($\boldsymbol{e}$) is the same as defined above, while for the additive genetic effect $\boldsymbol{a}$ the distribution is defined as $\boldsymbol{a} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{H}_S\sigma_a^2)$ where the structure of $\boldsymbol{H}_S$ is expressed by $\begin{bmatrix} \boldsymbol{TGT'} + \boldsymbol{D} & \boldsymbol{TG} & \boldsymbol{TZB} \\ \boldsymbol{GT'} & \boldsymbol{G} & \boldsymbol{ZB} \\ \boldsymbol{BZ'T'} & \boldsymbol{BZ'} & \boldsymbol{B} \end{bmatrix}$ with $\boldsymbol{D} = (\boldsymbol{A}^{11})^{-1}$, $\boldsymbol{T} = \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}$, and $\boldsymbol{B}$ representing a diagonal matrix of the form $\boldsymbol{I}\frac{1}{\sum_{i=1}^{N} 2p_i(1-p_i)}$ with p_i denoting the allele frequency of the i-th SNP, *N* is the number of SNPs, and $\boldsymbol{A}^{11}$ is the upper left block corresponding to $\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}^{-1}$. $\boldsymbol{X}$ and $\boldsymbol{W}_S$ denote the corresponding design matrices [4]. Note that the variance components and the proportion of residual additive polygenic variance were not estimated, instead their values corresponding to the parameters used in the Polish national genetic and genomic evaluation (Table 2).

Table 2. Variance components underlying the analyzed phenotypes

Trait	σ_a^2	σ_e^2	σ_u^2
Stature	5.50	4.63	20% σ_a^2
Foot angle	0.11	1.06	20% σ_a^2

Solving the GEBV prediction equations

Solutions for the effects fitted in the above models were obtained using the MiXBLUP 3.0 software [8] that solves the following equation: $\boldsymbol{D}^{-1}\boldsymbol{M}^{-1}\boldsymbol{C}\boldsymbol{x} = \boldsymbol{D}^{-1}\boldsymbol{M}^{-1}\boldsymbol{p}$, where $\boldsymbol{C}$ represents the coefficient matrix corresponding to the Mixed Model Equations (MME) for solving (1) or (2), $\boldsymbol{x}$ is the vector of model parameters, and $\boldsymbol{p}$ is the RHS of MME, while $\boldsymbol{M}$ and $\boldsymbol{D}$ respectively represent the first-level and the second-level preconditioning matrices. The computations were performed on a dedicated computing server running the Linux Red Hat operating system with 260GB of RAM, 16 Intel Xeon CPUs with 2.20GHz, and 600GB of hard disk space. For a large national dairy population, such as the one used in our study, the numerically severe challenge is to obtain the inverse of the $\boldsymbol{H}_G$ and $\boldsymbol{H}_S$ matrix, in particular of its component related to genotyped individuals – the dense sub-matrix $\boldsymbol{G}$. In our study, the following approaches were considered.

The **GT** approach [9], in which the $\boldsymbol{G}^{-1}$ matrix is represented by $\frac{1}{w}\boldsymbol{A}_{22}^{-1} - \frac{1}{w}\boldsymbol{T}'\boldsymbol{T}$, with $\boldsymbol{T} = \boldsymbol{L}^{-1}\boldsymbol{Z}'\boldsymbol{A}_{22}^{-1}$ where w denotes the proportion of a residual polygenic variance and $\boldsymbol{L}$ is defined by $\boldsymbol{Z}'\boldsymbol{A}_{22}^{-1}\boldsymbol{Z} + w\boldsymbol{I} = \boldsymbol{L}\boldsymbol{L}'$.

The **APY** approach [6] that divides the genotyped individuals into a core and non-core sub-groups for which the inverse is handled differently in such a way that the exact inverse is computed only for the core animal sub-group ($\boldsymbol{G}_c$) and the covariance between core and non-core individuals, while the part of the matrix corresponding to the non-core genotyped animals ($\boldsymbol{G}_n$) is handled as a diagonal matrix: $\boldsymbol{G}^{-1} \approx \begin{bmatrix} \boldsymbol{G}_c^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} + \begin{bmatrix} -\boldsymbol{G}_c^{-1}\boldsymbol{G}_{cn} \\ \boldsymbol{I} \end{bmatrix} \boldsymbol{M}_n^{-1} [-\boldsymbol{G}_{nc}\boldsymbol{G}_c^{-1} \quad \boldsymbol{I}]$ where the subscripts n and c represent non-core and core individuals respectively, and $\boldsymbol{M}_n$ is a diagonal matrix. In our study, four approaches to the selection of core animals were considered: APY3000top – where 3,000 genotyped bulls with the highest effective daughter contributions (EDC) were selected as core individuals, APY3000random – where the 3,000 core individuals were selected randomly from the genotyped population and the corresponding versions termed APY15000top, APY10000random, and APY15000random implementing 10,000 and 15,000 core individuals respectively. APY10000random proposed by Misztal et al. [10] formed a basis to asses differences between scenarios implementing a smaller (APY3000) as well as a larger (APY15000) selections of core animals.

The **ssSNP-BLUP** approach [4] does not meet the numerical burden of the models based on $\boldsymbol{G}$, since in terms of the modelling of genomic information only the inverse of the diagonal matrix $\boldsymbol{B}$ is required: $\boldsymbol{H}_S^{-1} = \begin{bmatrix} \boldsymbol{A}^{11} & \boldsymbol{A}^{12} & \boldsymbol{0} \\ \boldsymbol{A}^{21} & \boldsymbol{A}^{22} + \left(\frac{1}{w} - 1\right)\boldsymbol{A}_{22}^{-1} & -\frac{1}{w}\boldsymbol{A}_{22}^{-1}\boldsymbol{Z} \\ \boldsymbol{0} & -\frac{1}{w}\boldsymbol{Z}'\boldsymbol{A}_{22}^{-1} & \frac{1}{1-w}\boldsymbol{B}^{-1} + \frac{1}{w}\boldsymbol{Z}'\boldsymbol{A}_{22}^{-1}\boldsymbol{Z} \end{bmatrix}$, where $\begin{bmatrix} \boldsymbol{A}^{11} & \boldsymbol{A}^{12} \\ \boldsymbol{A}^{21} & \boldsymbol{A}^{22} \end{bmatrix}$ represent the blocks corresponding to $\begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix}^{-1}$.

Validation of GEBV

For the validation of GEBV prediction, bulls born after 2013 (7,296 bulls for stature; 6,200 bulls for foot angle) and cows born after 2016 (96,772 cows for stature; 96,771 cows for foot angle) were removed from the phenotype vector ($\boldsymbol{y}$) of equations (1) and (2). Based on such truncated data sets, the GEBVs of bulls with EDC greater than 20 were predicted by models (1) and (2). Prediction accuracy was assessed by a weighted linear regression [11]: $\boldsymbol{GEBV}_f = b_0 + b_1\boldsymbol{GEBV}_t + \boldsymbol{e}$, with $\boldsymbol{GEBV}_f$ representing the vector of GEBVs predicted based on the full data set with all available individuals while $\boldsymbol{GEBV}_t$ contains the GEBVs predicted based on the truncated data set. For i-th bull, weights were defined as $\frac{EDC_i}{EDC_i+k}$ with $k = \frac{4-h^2}{h^2}$. The unbiased evaluation is then represented bu b_1 equal to 1.00.

The above linear regression equation was fitted using the *lm* function and Pearson correlation coefficients, used to assess differences between GEBVs across scenarios were calculated using R software (www.rstudio.com).

Results

Model validation

The validation data set for stature and foot angle comprised 1,727 and 1,725 bulls with EDC > 20, respectively. For stature, generally, no marked differences in the estimated slope of the linear regression were observed between the models, varying between 0.94 (APY3000top) and 1.02 (APY10000random). Furthermore, except for APY3000, the differences between R^2 corresponding to each model were small. For stature, the GT approach achieved the highest R^2 of 0.83, while the lowest R^2 corresponded to 3,000 core animal scenarios with 0.57 for APY3000top and 0.60 for APY3000random. Similar results were observed for foot angle, albeit with overall lower R^2. GT and SNP-BLUP achieved the highest value of 0.76, while the lowest of 0.60 was attributed to APY3000random and 0.62 for APY3000top. Details of the validation results are summarized in (Table 3). Pearson's correlations (Fig. 1) between GEBVs predicted for the validation bulls from the full and the truncated datasets were calculated for all pairs of models. For stature, they demonstrated a good agreement only between SNP-BLUP and GT, as expressed by correlations of 0.996. Interestingly, the lowest correlations of 0.810 were observed between both models implementing APY3000 (i.e. APY3000random and APY3000top), as well as between APY3000top and APY15000top. For foot angle correlations are generally higher, and the pattern is very similar, with the correlation between SNP-BLUP and GT reaching 0.999 and the lowest correlation of 0.870 observed between both APY3000 scenarios, between GT and APY3000top, as well as between GP and APY3000random. Notably, for both traits, the correlations within APY3000top and within APY3000random are markedly lower than for the other models.

Table 3. Validation of GEBV predictions

Model variants	$\hat{b}_0$	SE($\hat{b}_0$)	$\hat{b}_1$	SE($\hat{b}_1$)	R^2
Stature, 1,727 validation bulls, h^2 = 0.54					
SNP-BLUP	-3.90	0.32	1.01	0.01	0.77
GT	-4.40	0.27	1.01	0.01	0.83
APY3000top	-1.81	0.44	0.94	0.02	0.57
APY3000random	-2.20	0.44	0.99	0.02	0.60
APY10000random	-3.74	0.36	1.02	0.02	0.72
APY15000top	-2.59	0.32	0.96	0.01	0.75
APY15000random	-2.32	0.33	0.96	0.01	0.73
Foot angle, 1,725 validation bulls, h^2 = 0.09					
SNP-BLUP	-2.03	0.18	1.03	0.01	0.76
GT	-1.96	0.18	1.04	0.01	0.76
APY3000top	-0.68	0.21	0.97	0.02	0.62
APY3000random	-0.52	0.26	1.03	0.02	0.60
APY10000random	-2.04	0.21	1.02	0.02	0.69
APY15000top	-2.15	0.18	1.02	0.01	0.75
APY15000random	-2.10	0.19	1.03	0.02	0.73

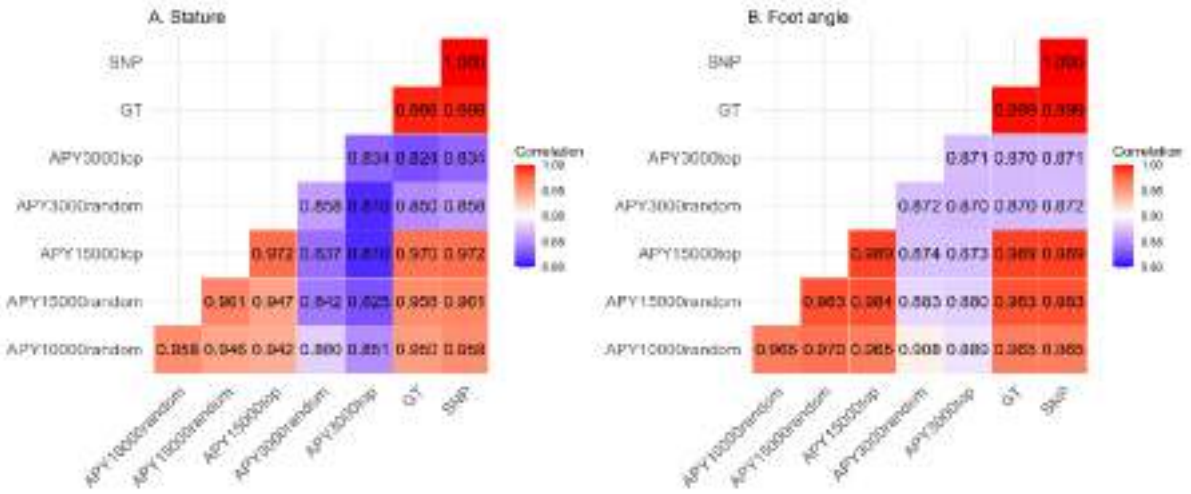


Figure 1. Pearson correlation coefficients of GEBVs predicted by different model variants

GEBV predictions

Figure 2 shows the differences between the GEBVs predicted by ssSNP-BLUP compared to the GEBVs predicted by the six ssG-BLUP implementations. For the difference in GEBV between SNP-BLUP and GT, all genotyped animals were considered, while for all comparisons involving APY, only the core individuals were used. The original solutions were rescaled for each model by subtracting the mean of a cow base population to provide GEBVs comparable across all models. For stature, the differences were generally small across bull birth years (2A), however, for foot angle a different pattern emerged (2B) as for all comparisons, except GT and the most informative APY15000top, the differences were unstable across bull birth years. Furthermore, we compared correlations of GEBVs predicted by the full and the truncated models for individuals representing base cows (i.e. cows with phenotype records) and bulls (i.e. bulls, which have a minimum of one daughter). The dashed line divides the plot into animals defined as old and young in the validation process. For stature, regardless of the model, correlations between full and truncated data sets, for old animals, were very close to unity. For young candidates, we observed a declining trend with the lowest correlations calculated for APY3000random (Fig. 3A). For foot angle, similar results were obtained, except for APY3000random for females, where we observed a decreasing correlation trend as early as from 2004. From 2008, correlations dropped below 0.9 (Fig. 3B). Finally, we compared the overlap of bulls with the top 50 GEBV predictions resulting from different models (Fig. 4). The number of bulls common across all models for stature (Fig. 4A) was 39, while the highest number of exclusive bulls (3) was observed in the top50 list resulting from the APY15000top. Regarding foot angle (Fig. 4B), the number of common bulls was even lower, being 31, while 8 were exclusive for the APY3000random.

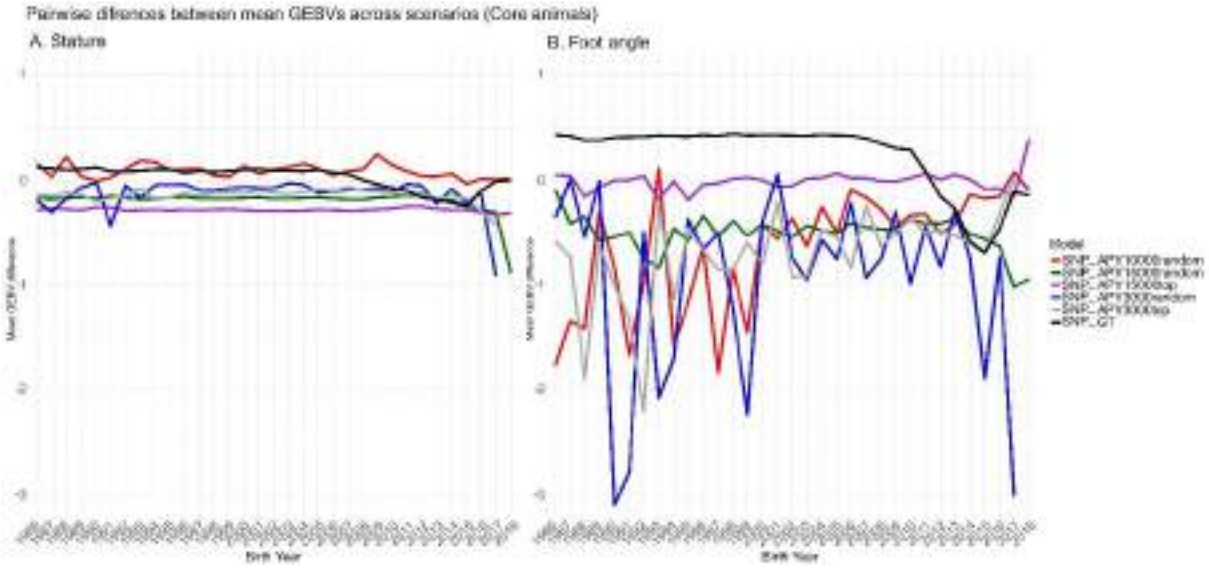


Figure 2. Mean GEBV difference divided by the genetic standard deviation between SNP-BLUP and GT, as well as SNP-BLUP and APY models implementing different scenarios of core individual selection. (A) stature, (B) foot angle

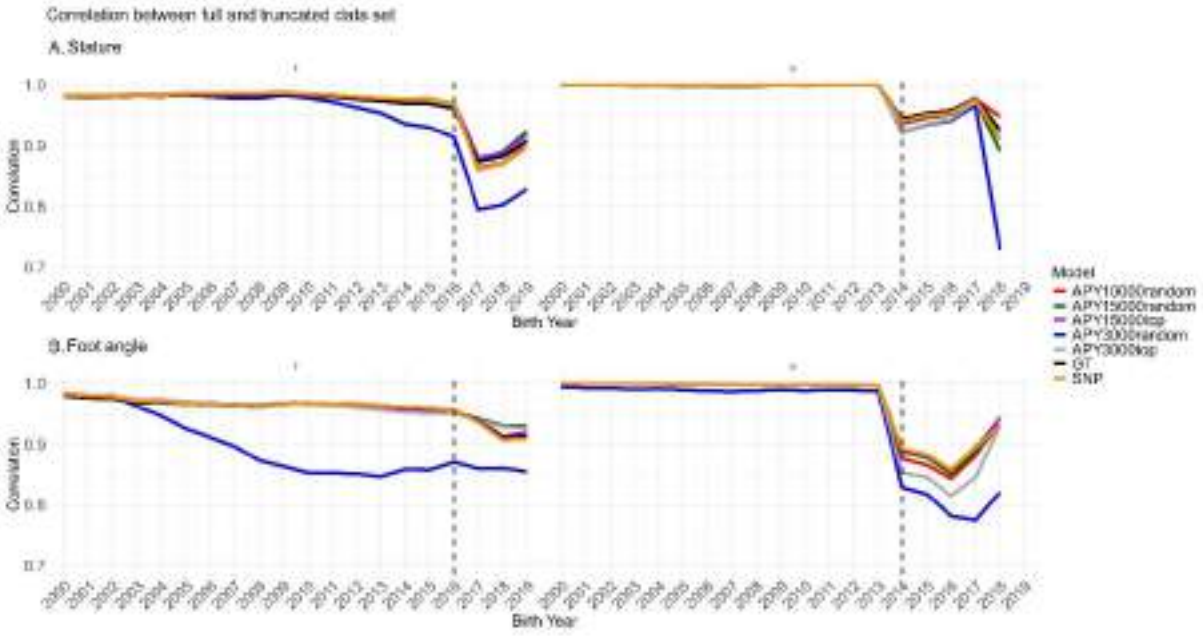


Figure 3. Correlation between full and truncated data set for cows with phenotype records and bulls with a minimum of one daughter. (A) stature, (B) foot angle

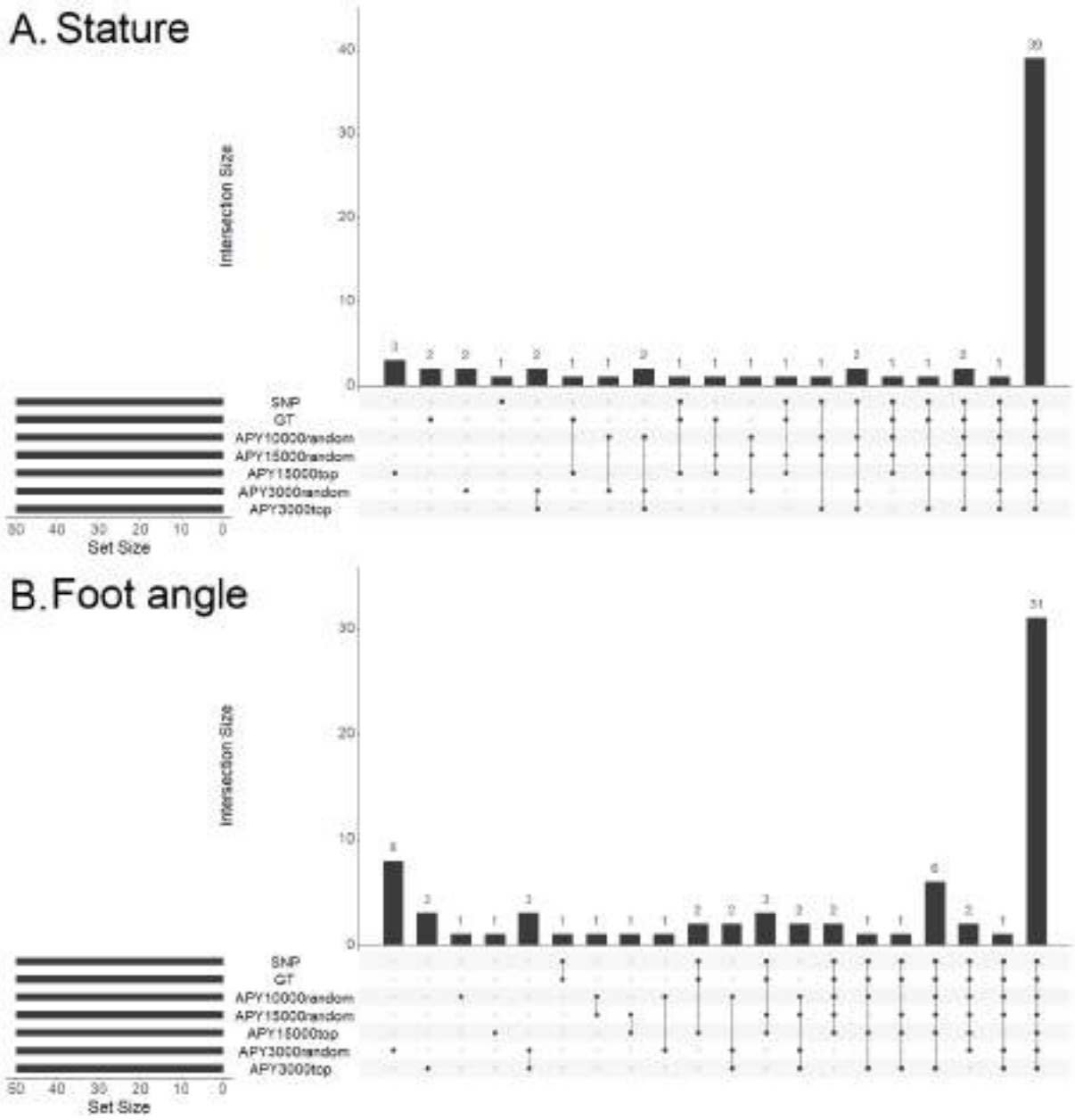


Figure 4. An upset plot of bulls with the top 50 GEBVs. (A) stature, (B) foot angle

Computational resources

The wall clock time corresponding to setting up and solving models (1) or (2) using the MiXBLUP software and the peak memory consumption varied considerably between solvers. The exact values are specified in (Table 4), but generalizing for both traits, SNP-BLUP and APY3000 were the fastest, closely followed by APY10000 random. The wall clock time of GT was the longest, twice the time of APY15000top. The peak memory consumption by SNP-BLUP was on the order of ten times lower than for the remaining solvers. In the case of iterations, SNP-BLUP required the most iterations (673 for stature and 1,027 for foot angle) and an average of 2.3 seconds per iteration. The lowest number of 335 iterations and an average

of 0.18 seconds per iteration were used for stature by APY3000top and 499 and 0.18 seconds for foot angle by APY3000random.

Table 4. Computational resources utilized by the solvers

Model variants	Wall clock time (min)		Peak RAM consumption (GB)		Number of iterations	
	Stature	Foot angle	Stature	Foot angle	Stature	Foot angle
SNP-BLUP	23	32	5.81	5.81	673	1027
GT	138	143	63.89	63.88	477	629
APY3000top	23	29	49.48	49.47	335	811
APY3000random	23	32	49.48	49.47	390	499
APY10000random	32	34	56.53	56.53	469	652
APY15000top	68	70	61.56	61.56	425	625
APY15000random	54	57	61.56	61.56	477	551

Discussion

Many studies have been conducted related to comparing single-step genomic prediction models with conventional two-step approaches. However, the literature on comparisons within the single-step modelling frame is scarce. Koivula et al. [12] considered genomic prediction models similar to (1) and (2). However, they addressed only the genotyped part of the population. Similarly to the results of our study, these authors observed very high correlations between models for predicted bull GEBVs or DGVs (direct genetic values) and similar validation results. Gao et al. also considered the validation performance of single-step models [13]. Although no marked differences were observed, the authors indicated that for APY-based solvers, not only the number but also the selection of core individuals was a crucial step influencing the prediction accuracy [13],[14],[15],[16]. Except worse prediction performance of APY3000random for cows, no marked differences due to the composition of the core animal set were observed in our study. Still, Macedo et al. [17] demonstrated low robustness in predictions with ssG-BLUP model towards differential handling of missing parent information. Moreover, our results demonstrated that a large number of core individuals is recommended, provided the availability of computational resources, especially RAM, to obtain stable prediction with ssG-BLUP. Similarly to the results of Misztal et al. [10], we also did not observe a marked gain in the prediction quality of GEBV when using more than 10,000 core individuals.

Regarding the computational aspects, we observed very large differences in RAM consumption between SNP-BLUP and other solvers. GT needed over 10 times more RAM and fewer iterations than SNP-BLUP, as was also demonstrated by Vandenplass et al. [18]. Two possible ways to circumvent the problem of the optimal choice of core individuals are either

the use of GT or SNP-BLUP solver-based prediction. The application of GT comes with the price of high memory requirements and long computing times. SNP-BLUP does not consume much memory and is computationally more efficient. Still the approaches did not yield identical rankings of top 50 bulls. Although 62% (foot angle) and 75% (stature) bulls were common in rankings across scenarios, there was still a considerable number of individuals that were ranked scenarios specific.

Conclusion

Regarding the prediction of GEBV on the active population scope, no marked differences between solvers (except the APY with only 3,000 core individuals) were observed, still GT and SNP-BLUP solvers resulted the highest prediction accuracy, while the effectiveness of the APY model depends on the size of the set of core animals: small cores lead to a noticeable decrease in accuracy, while cores with large numbers minimize these differences. Another factor influencing the choice of the solver is its computational efficiency expressed by the computation time and memory resources required, which was also indicated by Koivula et al. [9]. However, it should be kept in mind that the ranking of the top bulls is not identical between models, which has implications for the breeding industry in terms of semen pricing and selection since, typically, the top-ranking bulls are the most widely used.

Considering the abovementioned aspects, the results of our study indicate that GT and SNP-BLUP solvers offer more modelling and computational advantages.

CRediT authorship contribution statement

Dawid Słomian: Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft. **Joanna Szyda**: Conceptualization, Methodology, Project administration, Supervision, Writing – original draft.

Declaration of competing interest

The authors declare no competing interests.

Funding

This study was supported by a grant from the Ministry of Agriculture and Rural Development (DŻW.pp.862.1.2023).

References

Legarra A., Christensen O.F., Aguilar I., Misztal I. (2014). Single Step, a general approach for genomic selection. Livest. Sci. 166: 54–65.

Aguilar I., Misztal I., Johnson D.L., Legarra A., Tsuruta S., Lawlor T.J.(2010). Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93: 743–752.

Christensen O.F., Lund M.S. (2010). Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42: 2.

Liu Z., Goddard M.E., Reinhardt F., Reents R. (2014). A single-step genomic model with direct estimation of marker effects. J. Dairy Sci. 97: 5833–5850.

Liu Z., Goddard M.E., Hayes B.J., Reinhardt F., Reents R. (2016). Technical note: Equivalent genomic models with a residual polygenic effect. J. Dairy Sci. 99: 2016–2025.

Misztal I., Legarra A., Aguilar I. (2014). Using recursion to compute the inverse of the genomic relationship matrix. J. Dairy Sci. 97:3943–3952.

Lourenco D., Legarra A., Tsuruta S., Masuda Y., Aguilar I., Misztal I. (2020). Single-Step Genomic Evaluations from Theory to Practice: Using SNP Chips and Sequence Data in BLUPF90. Genes 11: 790.

Vandenplas J., Veerkamp R.F., Calus M.P.L., Lidauer M.H., Strandén I., Taskinen M., Schrauf M.F.S.G., ten Napel J. (2022). MiXBLUP 3.0 – software for large genomic evaluations in animal breeding programs. Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP), Rotterdam, the Netherlands: Wageningen Academic Publishers 1498–1501.

Mäntysaari E.A., Koivula M., Strandén I. (2020). Symposium review: Single-step genomic evaluations in dairy cattle. J. Dairy Sci. 103: 5314–5326.

Misztal I. (2016). Inexpensive Computation of the Inverse of the Genomic Relationship Matrix in Populations with Small Effective Population Size. Genetics 202: 401–409.

Mäntysaari E.A., Liu Z., VanRaden P. (2010) Interbull validation test for genomic evaluations. Interbull Bulletin 17.

Koivula M., Strandén I., Su G., Mäntysaari E.A. (2012). Different methods to calculate genomic predictions—Comparisons of BLUP at the single nucleotide polymorphism level (SNP-BLUP), BLUP at the individual level (G-BLUP), and the one-step approach (H-BLUP). J. Dairy Sci. 95: 4065–4073.

Gao H., Koivula M., Jensen J., Strandén I., Madsen P., Pitkänen T., Aamand G.P, Mäntysaari E.A. (2018). Short communication: Genomic prediction using different single-step methods in the Finnish red dairy cattle population. J. Dairy Sci. 101: 10082–10088.

Fragomeni B.O., Lourenco D.A.L., Tsuruta S., Masuda Y., Aguilar I., Legarra A., Lawlor T.J., Misztal I. (2015). Hot topic: Use of genomic recursions in single-step genomic best linear unbiased predictor (BLUP) with a large number of genotypes. J. Dairy Sci. 98: 4090–4094.

Masuda Y., Misztal I., Tsuruta S., Legarra A., Aguilar I., Lourenco D.A.L., Fragomeni B.O., Lawlor T.J. (2016). Implementation of genomic recursions in single-step genomic best linear unbiased predictor for US Holsteins with a large number of genotyped animals. J. Dairy Sci. 99: 1968–1974.

Strandén I., Garrick D.J. (2009). Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J. Dairy Sci. 92: 2971–2975.

Macedo F.L., Astruc J.M., Meuwissen T.H.E., Legarra A. (2022). Removing data and using metafounders alleviates biases for all traits in Lacaune dairy sheep predictions. J. Dairy Sci. 105: 2439–2452.

Vandenplas J., ten Napel J., Darbaghshahi S.N., Evans R., Calus M.P.L., Veerkamp R., Cromie A., Mäntysaari E.A., Strandén I. (2023). Efficient large-scale single-step evaluations and indirect genomic prediction of genotyped selection candidates. Genet. Sel. Evol. 55: 37.

Received: 15 II 2025
Accepted: 8 VIII 2025

Comparison of BLUPF90IOD3 and MiXBLUP implementations of the single-step model applied to the Polish national dairy cattle evaluation

Dawid Słomian
dawid.slomian11@gmail.com
National Research Institute of Animal Production https://orcid.org/0000-0002-9037-7703

Michalina Jakimowicz
michalina.jakimowicz@upwr.edu.pl
Wroclaw Univeristy of Environmental and Life Sciences

Tomasz Suchocki
tomasz.suchocki@upwr.edu.pl
Wroclaw Univeristy of Environmental and Life Sciences

Joanna Szyda
joanna.szyda@upwr.edu.pl
Wroclaw Univeristy of Environmental and Life Sciences

Research Article

Keywords: BLUPF90IOD3, GEBV, G-BLUP, MiXBLUP, single-step

DOI: https://doi.org/10.21203/rs.3.rs-8398690/v1

Additional Declarations: The authors declare no competing interests.

Comparison of BLUPF90IOD3 and MiXBLUP implementations of the single-step model applied to the Polish national dairy cattle evaluation

Dawid Słomian[a], Michalina Jakimowicz[b], Tomasz Suchocki[a,b], and Joanna Szyda[a,b]

[a] National Research Institute of Animal Production, Krakowska 1, 32-083 Balice, Poland

[b] Wroclaw University of Environmental and Life Sciences, Department of Genetics, The Biostatistic Group, Kozuchowska 7, 51-631 Wroclaw, Poland

Corresponding author

Joanna Szyda, joanna.szyda@upwr.edu.pl

Abstract

The integration of phenotypic, genomic, and pedigree data into a single-step model for predicting genomically enhanced estimated breeding values (GEBVs) has become crucial for the accurate genetic evaluation of dairy cattle. This study compared two widely used software implementations, MiXBLUP and BLUPF90IOD3, for the prediction of breeding values using the single-step G-BLUP model based on data from the Polish national evaluation for stature. Four core animal sets were tested, which differed in the selection of bulls and cows. The GEBVs were predicted and validated using different subsets of the population. Both software packages resulted in high correlations (0.89 and 0.97) between full and truncated dataset predictions and similar validation performance, with MiXBLUP exhibiting slightly greater consistency across different sets of core animals. The ranking of the top 50 bulls remained stable across the implementations. This study concludes that both software implementations provide comparable GEBV predictions, suggesting that software choice should consider computational efficiency, cost, and modeling flexibility, with MiXBLUP offering additional options for GEBV estimation.

Keywords

BLUPF90IOD3, GEBV, G-BLUP, MiXBLUP, single-step

INTRODUCTION

For several years, the single-step approach to predict breeding values has become increasingly popular, and in many countries, work is underway to implement it into the routine evaluation system for dairy cattle. The growing importance of the single-step model is due to the possibility of integrating phenotypic, genomic, and pedigree data, which results in the prediction of breeding values for all individuals under one unified model without the need to conduct two separate evaluations (i.e., a conventional and a genomic).. In national genomic evaluations of dairy cattle, two forms of a single-step model are used – the single-step G-BLUP fitting a random animal additive genetic effect with a relationship matrix defined by the pedigree and/or SNP genotype information (Aguilar et al., 2010; Christensen & Lund, 2010) and the single-step SNP-BLUP fitting both a random animal additive effect defined above and a random SNP effect (Liu et al., 2014).

The primary purpose of our study was to compare the breeding values from the G-BLUP model predicted by two software implementations that are most widely used on a national scale, MiXBLUP (Vandenplas et al., 2022) and BLUPF90IOD3 (Aguilar et al., 2018) using the same model parametrization and a dataset. This was done by considering various sets of core animals using data from the Polish national evaluation for stature.

MATERIALS AND METHODS

Data

The analyzed data represent the active population of animals that entered the Polish national genetic evaluation for stature ($h^2 = 0.54$) from December 2021. It includes 1,098,611 cow phenotypes and 141,397 pseudophenotypes expressed by deregressed proofs (DRP) from the multiple across-country evaluation (MACE) carried out by Interbull. DRPs were adjusted for the phenotypes of bulls' daughters born in Poland. Most genotyped individuals were genotyped using various versions of the EuroG MDIllumina genotyping microarray, which was custom-designed for the EuroGenomics Cooperative. Individuals genotyped with other commercial

platforms were imputed to EuroG MD using the Fimpute software (Sargolzaei et al., 2014). The SNP preselection criteria followed the procedure used in the national genomic evaluation in Poland. The criteria comprised a minor allele frequency of at least 0.01 and a technical quality of genotyping expressed by a minimum call rate of 99%. After editing, 46,118 SNPs remained for further analysis. The genomic data contained 42,134 cow genotypes and 47,108 bull genotypes. Full pedigree information was truncated after the fifth generation using the Relax2 software (Stranden and Vuori, 2006) prune 5 option. It resulted in 1,555,995 individuals and 33 Unknown Parent Groups (UPGs) based on birth year, country of origin, and sex.

Prediction of breeding values

The prediction of genomically enhanced breeding values (GEBV) was based on the following single-step G-BLUP model:

$$y = Xb + Wa + e\ , \quad (1)$$

where y is the vector of dependent variables represented by cows' measured phenotypes for stature and bulls' pseudophenotypes expressed by their MACE DRPs (Jairath et al., 1998), b represents a vector of fixed effects including age at calving, lactation phase, and herd corresponding to cows' phenotypes as well as corresponding phantom codes of the fixed effects for bulls' DRPs, a represents a vector of breeding values, and e is the vector of residuals. The underlying covariance structure of the random effects is given by $a \sim N\left(0, H_G \sigma_a^2\right)$ and $e \sim N\left(0, R\sigma_e^2\right)$. H_G is given by $\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ $+ \begin{bmatrix} A_{12}A_{22}^{-1}(G - A_{22})A_{22}^{-1}A_{21} & A_{12}A_{22}^{-1}(G - A_{22}) \\ (G - A_{22})A_{22}^{-1}A_{21} & G - A_{22} \end{bmatrix}$ (Lourenco et al., 2020), where A_{11}, A_{12}/A_{21}, and A_{22} are the components of the numerator relationship matrix constructed based on the pedigree corresponding to non-genotyped

animals, the covariance between non-genotyped and genotyped animals, as well as between genotyped animals, respectively, while G represents the genomic relationship matrix between genotyped animals. R is a diagonal matrix containing 1.00 for cows with phenotypes or n_i for bulls with MACE DRPs, with n_i representing a difference in effective daughter contributions of i-th bull between the MACE and the national evaluation. X and W denote the corresponding design matrices. For solving the mixed model equations corresponding to model (1) an inverse of the genomic covariance matrix (G) is required. Following (Misztal, 2016) the inverse was approximated as: $G^{-1} \approx \begin{bmatrix} G_{cc}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -G_{cc}^{-1}G_{cn} \\ I \end{bmatrix} M_{nn}^{-1} \begin{bmatrix} -G_{cn}^{T}G_{cc}^{-1} & I \end{bmatrix}$, where G_{cc} represents the genomic relationship matrix for the subgroup of animals defined as *core* individuals, G_{cn} is the genomic relationship matrix between *core* and *non-core* individuals, and M_{nn} is a diagonal matrix with nonzero elements corresponding to the variance of the mendelian sampling effect for each *non-core* individual. Four sets of genotyped animals were used as the core individuals. The **All_male** set was composed of all bulls with phenotypes, the **Male_20K** set was composed of 20,000 bulls randomly selected from the active population, the **Female_30K** set was composed of 30,000 cows randomly selected from the active population, and the **Random_20K** set was composed of 20,000 individuals (bulls and cows) randomly selected from the active population. The random choice of core animals was performed using a custom-written R script using the *sample* function. The first three sets were chosen to represent markedly different scenarios that would allow for a better understanding of the impact of the selected core

117 animals for prediction. The random selection resembles the scenario used
118 in practical applications.

119 The computations were performed using two software packages,
120 MiXBLUP and BLUPF90IOD3, which implement the PCG solver with an
121 equivalent convergence criterion (Vandenplas et al., 2021; Masuda, 2019).
122 The corresponding and equivalent stopping criteria were given by 1E-07 for
123 MiXBLUP and 1E-14 for BLUPF90IOD3. The difference between programs
124 was due to because of the differences in software implementation.

125 ***Validation of predictions***

126 The validation of GEBV prediction followed the GEBVtest method
127 (Mäntysaari et al., 2010) adopted by the Interbull organization
128 (www.interbull.org) and was based on the following linear regression model:

129 $$GEBV_{full} = b_0 + b_1 GEBV_{trunc} + \varepsilon, \quad (2)$$

130 where $GEBV_{full}$ is the vector of GEBV predicted by equation (1) for the
131 complete data set and $GEBV_{trunc}$ represents a vector of GEBVs predicted by
132 equation (1) for the validation dataset. Three validation sets were defined**:**
133 **all_validation** – composed of all bulls born between 2014 and 2017 with
134 Effected Daughter Contributions (EDC) over 20; **daughters_in_PL** – a
135 subset of **all_validation** set composed only of bulls with daughters in
136 Poland; **young** – comprising genotyped young bulls born after 2017.

137 The validation model (2) was fitted using the *lm* function in the R
138 software (R Core Team, 2021).

RESULTS

The correlation coefficients between the GEBVs predicted by the full and each of the three validation datasets were very similar across software implementations, the core animal, and the validation sets (Table 1). They varied between 0.89 for the scenario with **Female_30K** and **Daughters_in_PL** set predicted by BLUPF90IOD3 and 0.97 for the scenario with **Male_20K** and **young** set, regardless of the software. Considering validation, the estimated regression intercepts $\left(\hat{\beta}_0\right)$ varied not only between software implementations, but also between validation scenarios, and core animal sets. Albeit similarities of intercepts were not expected, as the various subsets of data were not corrected for the genetic base. Moreover, from the validation perspective, the most important quantities are $\hat{\beta}_1$ and R^2 that remained stable across software implementations for each validation scenario (Table 2-Table 4). Note, that slope estimates obtained for the **young** dataset were very close to unity, which is due to the fact that young bulls have no daughter information in both, the full as well as in the truncated set. For MiXBLUP, the number of iterations to convergence ranged between 364 for **Random_20K**, truncated dataset, and 551 for **Female_30K**, full dataset, and for each scenario, the full dataset required more iterations than the truncated dataset. For most scenarios, the convergence of BLUPF90IOD3 required more iterations than MiXBLUP and was more variable across the datasets. The lowest number of 284 iterations was needed to converge the **Random_20K** truncated data set, while 707 iterations were required to converge for the full version of this scenario.

Surprisingly, the smaller, i.e., truncated, dataset does not always result in faster convergence of the BLUPF90IOD3 solver. A summary of the number of iterations required for each scenario is presented in Table 5. Further similarities between the programs extend to their high robustness towards different sets of core animals (Figure 1), since among the top 50 ranked bulls, 45 and 44 were also present in the top 50 group defined by the other three scenarios for MiXBLUP and BLUPF90IOD3 implementations, respectively. Within each scenario, there were almost no differences in bull rankings, so the numbers of common bulls were 49 and 48.

DISCUSSION

This study aimed to compare two software implementations of the G-BLUP model, most commonly used for routine genetic evaluations on a national scale, MiXBLUP and BLUPF90IOD3. Both programs provided very similar performance in terms of the quality of prediction of GEBVs, expressed either by the correlation between GEBVS predicted from truncated and full datasets or by the official criterion applied to national routine evaluations, expressed by the Interbull GEBV validation test. A comparison of different core animal scenarios showed a high compatibility of breeding value estimates. These results are consistent with the APY solver, which assumes that the number of individuals, rather than their detailed composition, is crucial (Misztal et al., 2014; Fragomeni et al., 2015). Studies indicate that the random selection of a large group of core animals provides accuracy comparable to cores created according to more advanced criteria (Misztal et al., 2020), which explains the similar results in our scenarios. Any slight

difference between the two software programmes is due to numerical differences in the algorithms rather than substantive errors. Practically, this means that both MiXBLUP and BLUPF90IOD3 can be used interchangeably for routine evaluation, providing reliable results. Finally, our results confirm that, regardless of the choice of software, one obtains reliable predictions of estimated breeding values. Both programs fulfill their role and provide similar results, while the main difference is in the usability and technical aspects rather than in the quality of the results. Regarding the availability of fitting different genetic evaluation models, MiXBLUP, apart from the G-BLUP, also allows for the implementation of the GT-BLUP and the SNP-BLUP solvers, which makes it more flexible in terms of the choice of GEBV estimation procedures.

In conclusion, our comparison demonstrated that both G-BLUP implementations are very similar in terms of predicted GEBVs, therefore, the choice of a particular program for routine national genetic evaluation needs to be based on other criteria, such as price, support, computational efficiency, or data modelling flexibility.

COMPETING INTERESTS

The author(s) declare none.

ACKNOWLEDGEMENTS

The computations were carried out on the server of the Center of Genetics of the Polish Federation of Cattle Breeders and Dairy Farmers.

REFERENCES

1. Aguilar, I., Tsuruta, S., Masuda, Y., Lourenco, D., & Misztal, I. (2018). BLUPF90 suite of programs for animal breeding with focus on genomics. *Proceedings of the 11th World Congress on Genetics Applied to Livestock Production, 751*
2. Aguilar, I., Misztal, I., Johnson, D., Legarra, A., Tsuruta, S., & Lawlor, T. (2010). Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *Journal of Dairy Science*, *93*(2), 743–752. https://doi.org/10.3168/jds.2009-2730
3. Christensen, O. F., & Lund, M. S. (2010). Genomic prediction when some animals are not genotyped. *Genetics Selection Evolution*, *42*(1). https://doi.org/10.1186/1297-9686-42-2
4. Fragomeni, B., Lourenco, D., Tsuruta, S., Masuda, Y., Aguilar, I., Legarra, A., Lawlor, T., & Misztal, I. (2015). Hot topic: Use of genomic recursions in single-step genomic best linear unbiased predictor (BLUP) with a large number of genotypes. *Journal of Dairy Science*, *98*(6), 4090–4094. https://doi.org/10.3168/jds.2014-9125
5. Jairath, L., Dekkers, J., Schaeffer, L., Liu, Z., Burnside, E., & Kolstad, B. (1998). Genetic evaluation for herd life in Canada. *Journal of Dairy Science*, *81*(2), 550–562. https://doi.org/10.3168/jds.s0022-0302(98)75607-3
6. Liu, Z., Goddard, M., Reinhardt, F., & Reents, R. (2014). A single-step genomic model with direct estimation of marker effects. *Journal of Dairy Science*, *97*(9), 5833–5850. https://doi.org/10.3168/jds.2014-7924

7. Lourenco, D., Legarra, A., Tsuruta, S., Masuda, Y., Aguilar, I., & Misztal, I. (2020). Single-Step Genomic Evaluations from Theory to Practice: Using SNP Chips and Sequence Data in BLUPF90. *Genes*, *11*(7), 790. https://doi.org/10.3390/genes11070790
8. Masuda, Y. (2019). Introduction to BLUPF90 suite programs. *Department of Animal and Dairy Science, University of Georgia.* https://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=tutorial_blupf90.pdf
9. Mäntysaari, E. A., Liu, Z., & VanRaden, P. (2010). Interbull validation test for genomic evaluations. Interbull Bulletin, 41, Article 41
10. Misztal, I. (2015). Inexpensive Computation of the Inverse of the Genomic Relationship Matrix in Populations with Small Effective Population Size. *Genetics*, *202*(2), 401–409. https://doi.org/10.1534/genetics.115.182089
11. Misztal, I., Legarra, A., & Aguilar, I. (2014). Using recursion to compute the inverse of the genomic relationship matrix. *Journal of Dairy Science*, *97*(6), 3943–3952. https://doi.org/10.3168/jds.2013-7752
12. Misztal, I., Lourenco, D., & Legarra, A. (2020). Current status of genomic evaluation. *Journal of Animal Science*, 98(4). https://doi.org/10.1093/jas/skaa101
13. Stranden, I., & Vuori, K. (2006). RelaX2: pedigree analysis program. Proceedings of 8th World Congress on Genetics Applied to Livestock Production, 27–30

14. Vandenplas, J., Calus, M. P. L., Eding, H., Van Pelt, M., Bergsma, R., & Vuik, C. (2021). Convergence behavior of single-step GBLUP and SNPBLUP for different termination criteria. *Genetics Selection Evolution*, *53*(1). https://doi.org/10.1186/s12711-021-00626-1

15. Vandenplas, J., Veerkamp, R., Calus, M., Lidauer, M., Strandén, I., Taskinen, M., Schrauf, M., & Napel, J. T. (2022). 358. MiXBLUP 3.0 – software for large genomic evaluations in animal breeding programs. Proceedings of 12th World Congress on Genetics Applied to Livestock Production, 1498–1501. https://doi.org/10.3920/978-90-8686-940-4_358

LIST OF TABLES AND FIGURES

Figure 1. Venn plot of top the top 50 GEBV bulls for different core animal sets within each software.

288 **Table *1***

	All validation 2,975 bulls		Daughters in PL 711 bulls		Young 1,737 bulls	
	MiXBLUP	**BLUPF90IOD3**	**MiXBLUP**	**BLUPF90IOD3**	**MiXBLUP**	**BLUPF90IOD3**
Male_20K	0.92 ± 0.002	0.92 ± 0.002	0.90 ± 0.005	0.90 ± 0.005	0.97 ± 0.001	0.97 ± 0.001
Random_20K	0.92 ± 0.002	0.92 ± 0.002	0.90 ± 0.005	0.90 ± 0.005	0.96 ± 0.001	0.95 ± 0.002
Female_30K	0.92 ± 0.002	0.92 ± 0.002	0.90 ± 0.005	0.89 ± 0.006	0.95 ± 0.002	0.94 ± 0.002
All_male	0.93 ± 0.002	0.93 ± 0.002	0.90 ± 0.005	0.90 ± 0.005	0.96 ± 0.001	0.96 ± 0.001

289

290

Table 2

2,975 validation bulls	$\hat{\beta}_0$	$\hat{\beta}_1$	R^2
	MiXBLUP		
Male_20K	4.702	0.844	0.851
Random_20K	5.212	0.829	0.849
Female_30K	5.442	0.818	0.846
All_male	4.608	0.856	0.858
	BLUPF90IOD3		
Male_20K	6.954	0.855	0.851
Random_20K	5.443	0.845	0.850
Female_30K	1.924	0.838	0.831
All_male	16.345	0.867	0.860

$\hat{\beta}_0$ – intercept

$\hat{\beta}_1$ – slope

R^2 – coefficient of determination

Table 3

711 validation bulls	$\hat{\beta}_0$	$\hat{\beta}_1$	R^2
MiXBLUP			
Male_20K	4.677	0.843	0.805
Random_20K	5.175	0.830	0.815
Female_30K	5.205	0.830	0.815
All_male	4.756	0.848	0.808
BLUPF90IOD3			
Male_20K	6.861	0.861	0.800
Random_20K	5.378	0.848	0.814
Female_30K	1.615	0.831	0.808
All_male	16.504	0.860	0.808

$\hat{\beta}_0$ – intercept

$\hat{\beta}_1$ – slope

R^2 – coefficient of determination

Table 4

1,737 validation bulls	$\hat{\beta}_0$	$\hat{\beta}_1$	R^2
MiXBLUP			
Male_20K	2.356	0.968	0.938
Random_20K	2.273	1.019	0.913
Female_30K	1.877	1.030	0.903
All_male	3.201	0.959	0.930
BLUPF90IOD3			
Male_20K	4.828	0.980	0.940
Random_20K	2.656	1.035	0.910
Female_30K	9.839	0.988	0.907
All_male	16.175	0.977	0.931

$\hat{\beta}_0$ – intercept

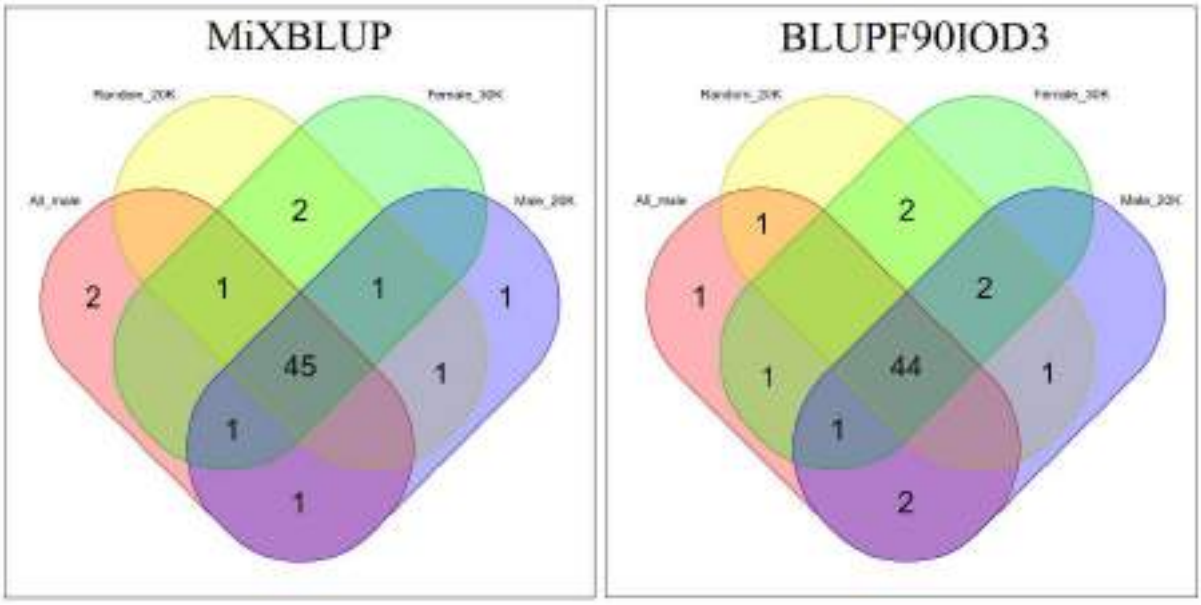
$\hat{\beta}_1$ – slope

R^2 – coefficient of determination

Table 5

	Number of iterations	
	MiXBLUP	**BLUPF90IOD3**
Male_20K	438	707
Random_20K	404	435
Female_30K	551	392
All_male	424	459
Male_20K_truncated	426	412
Random_20K_truncated	364	284
Female_30K_truncated	432	531
All_male_truncated	381	549

Figure 1



Modeling missing parents in single-step test-day SNP-BLUP evaluation of dairy cattle

Dawid Słomian[1], Jeremie Vandenplas[3], Jan Ten Napel[3], Kacper Żukowski[1], Monika Skarwecka[1], Joanna Szyda[1,2]

[1]National Research Institute of Animal Production, Krakowska 1, 32-083 Balice, Poland

[2]Biostatistic Group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, Kozuchowska 7, 51-631 Wroclaw, Poland

[3]Animal Breeding and Genomics, Wageningen University & Research, P. O. Box 338, 6700 AH Wageningen, Netherlands

Corresponding author: Joanna Szyda joanna.szyda@upwr.edu.pl

Abstract

In many countries, single-step genomic models have replaced multiple step models for routine evaluation. These models use all available information on animals' phenotypes, genotypes, and pedigrees, yet missing parental information in pedigrees remains a challenge that affects genomic breeding value (GEBV) predictions. Therefore, the choice of method for handling missing parents can affect the prediction of breeding values. Here, we compared three approaches to model missing parental information for three levels of missing pedigree data: P_Real – pedigree from routine evaluation, P_2010 – at least 20 percent of dams and 10 percent of sires born before 2019 were set to missing, and P_4020 – at least 40 percent of dams and 20 percent of sires born before 2019 were set to missing. Missing parents' information was expressed through missing codes in the raw pedigree (RP) by defining genetic groups (GG) that represent missing parents grouped based on year of birth, sex, and country of origin, or by defining metafounders (MF), which represent missing parents grouped by average genetic relationships estimated from the genomic information of their descendants. The genomic breeding values for fat yield were estimated using the single-step test-day SNP-BLUP model implemented with MiXBLUP software. For the considered scenarios, the results were presented separately for sires and dams, as well as for genotyped and ungenotyped individuals. We observed differences in the prediction quality between genotyped and ungenotyped animals. While GEBV predictions for the former were generally stable across scenarios, the predictions for the ungenotyped individuals varied. In particular, the removal of parental information led to less stable results when missing parental information was expressed by MF, where insufficient pedigree completeness resulted in an overestimation of the genetic trend. In conclusion, for informative pedigrees with a small percentage of missing parents, the incorporation of GG and MF results in very similar GEBV

36 predictions, however GG appear to be a more robust approach for ungenotyped individuals in

37 highly incomplete pedigrees.

38

Introduction

Currently, many countries implement a single-step model that includes all available animal information (phenotype, genotype, and pedigree) for routine genetic evaluation. The structure of pedigree data is an important aspect of the genetic evaluation of dairy cattle (Bradford et al., 2019). Therefore, a practical challenge is to deal with missing individuals in the pedigree. The standard approach for this situation is to combine missing individuals within unrelated genetic groups (or phantom parent groups), which are typically defined based on the country of origin, sex, and year of birth (Westell et al., 1988; Legarra et al., 2007). Although genetic groups are commonly used, they can lead to bias and excessive scattering of genomically estimated breeding values (GEBV), especially if there are many missing parents (Masuda et al., 2021; Himmelbauer et al., 2024). Defining metafounders (Legarra et al., 2015) is a genotype-aware alternative to genetic groups. This approach involves forming groups of missing parents based on average genetic relationships estimated from single nucleotide polymorphisms (SNPs) information of their descendants, for which a 0.5 allele frequency at all SNPs is assumed. However, Kudinov et al. (2020) did not observe any differences in GEBVs predicted using genetic groups and metafounder approaches to combine missing parents. On the other hand Bradford et al. (2019), Macedo et al. (2020), Kudinov et al. (2022), and Himmelbauer et al. (2024) reported more accurate predictions of breeding values when using metafounders as compared to genetic groups, especially in pedigrees with low pedigree completeness and for low heritable traits.

As all the above-mentioned studies compared the use of genetic groups and metafounders for GEBV prediction based on the G-BLUP model, we aimed to compare the approaches of handling missing pedigree in the single-step SNP-BLUP formulation. In particular, three levels of missing parent data and three approaches for handling them were considered. The comparison was applied to a real dataset representing the national evaluation for fat yield. In

particular, we focused on the GEBV validation results, GEBV trends, and accuracy of GEBV prediction for young individuals.

Material

The data originated from the Polish national evaluation run for fat yield from April 2024 (Table 1) and included 3,707,727 cows with 63,615,019 records in the full dataset and 3,224,917 cows with 58,446,695 records in the truncated dataset, defined by removing animals born after 2018. Genotypic information was available for 181,991 individuals and comprised 46,118 SNP genotypes. The pedigree used in this study contained 4,712,143 animals and originated from the raw pedigree file by extracting individuals representing the third generation before the oldest individual with data (phenotype or genotype) using the RelaX2 software (Strandén, 2014).

Based on this pedigree, the following scenarios were considered:

- Pedigree_real (**P_Real**) – the original pedigree from routine evaluation containing 262,519 (5.6%) missing sires and 719,360 (15.3%) missing dams,
- Pedigree_20_10 (**P_2010**) – **P_real** with approximately 20% of the dams and 10% of the sires set to missing, containing 446,669 (9.5%) missing bulls and 1,076,127 (22.8%) missing cows,
- Pedigree_40_20 **(P_4020**) – **P_real** with approximately 40% of the dams and 20% of the sires set to missing, containing 884,192 (18.7%) missing bulls and 1,868,957 (39.6%) missing cows.

Note that the pedigree of the youngest animals, born from 2019, that were used for validation was the same in all scenarios. Differences between scenarios, therefore, start with the pedigree of the second generation, that is, the parents of these young individuals (Figure 1).

Moreover, only parental information was removed, so that the phenotypic records of the parents remained in the data.

Furthermore, for the defined pedigrees, three approaches for handling missing parental information were implemented:

- Raw pedigree (**RP**) – missing parents' IDs set to missing,
- Genetic groups (**GG**) – missing parents replaced by unrelated **GG** defined based on birth year (10-year grouping starting from 1960, with individuals born before 1960 assigned to a common genetic group), country of origin (Poland, USA, Canada, other), and sex,
- Metafounders (**MF**) – missing parents replaced by **MF**, which represent genetic groups with relationships estimated from the genomic information of descendants.

The number of defined **GGs** differed between pedigree scenarios. For **P_2010** and **P_4020,** more **GGs** were defined than for **P_Real**. The highest rate of missing parents in each scenario was observed for animals born in Poland between 1990 and 2019 (Figure 2). Since **MF** were created based on **GG**, the numbers of **MF** and **GG** were the same.

Methods

The following single-step test-day SNP-BLUP model (Liu et al., 2004) was used to predict breeding values:

$$\boldsymbol{y} = \boldsymbol{Xh} + \boldsymbol{Wf} + \boldsymbol{Vp} + \boldsymbol{Vu} + \boldsymbol{e},$$

where $\boldsymbol{y}$ contains cow test day records for fat yield from the first three parities, $\boldsymbol{h}$ is a vector of fixed effects of herd-test_date-parity-milking_frequency, $\boldsymbol{f}$ is a vector of fixed lactation curve coefficients, which was modeled by the Wilmink function (Liu et al., 2004), $\boldsymbol{p}$ is a vector of permanent environmental effects expressed as three random regression coefficients of the Legendre polynomial, and $\boldsymbol{u}$ is a random additive genetic effect also described by the three

random regression coefficients of the Legendre polynomials. $\boldsymbol{V}$ is a design matrix for $\boldsymbol{u}$, $\boldsymbol{p}$ contains the Legendre coefficients for the first three lactations, $\boldsymbol{X}$ is the design matrix for the fixed herd-test-date-parity-milking-frequency effect $\boldsymbol{h}$, and $\boldsymbol{W}$ is the design matrix for fixed effect $\boldsymbol{f}$, and $\boldsymbol{e}$ is a residual. The model was implemented using MiXBLUP 3.0 (Vandenplas et al., 2022).

The validation of $\widehat{\boldsymbol{u}}$ was conducted on $\widehat{\boldsymbol{u}}$ representing the 305-day Genomically Enhanced Breeding Values (GEBVs) combined over all three lactations, using the following pattern $\boldsymbol{GEBV}t = 0.5\boldsymbol{GEBV}1 + 0.3\boldsymbol{GEBV}2 + 0.2\boldsymbol{GEBV}3$, where $\boldsymbol{GEBV}t$ is the combined GEBV and $\boldsymbol{GEBV}i$ represents GEBV for the i-th lactation. The validation test conducted following Mäntysaari et al. (2010) was implemented separately for cows and bulls based on the standardized $\boldsymbol{GEBV}t$ as:

$$\boldsymbol{SGEBV}t\_f = b_0 + b_1\boldsymbol{SGEBV}t\_p + \boldsymbol{e},$$

where $\boldsymbol{SGEBV}t\_f$ represents the vector of standardized GEBVs predicted based on the full dataset, while $\boldsymbol{SGEBV}t\_p$ represents standardized GEBVs predicted based on the truncated dataset, b_0 represents the intercept, which indicates a systematic bias in the model's prediction, and b_1 represents the regression slope, that is, the divergence between predictions and actual GEBVs. The R^2 coefficient that quantifies the percentage of variance of $\boldsymbol{SGEBV}t\_f$ explained by $\boldsymbol{SGEBV}t\_p$ was used as a measure of prediction accuracy. Validation cows were defined as dams born between 2019 and 2022 with test-day records for at least one lactation. Validation bulls were defined as sires born between 2017 and 2020 with at least 20 daughters. The standardization of GEBVs was performed separately for the full and truncated datasets as: $SGEBV_{ti} = \frac{GEBV_{ti} - mean(\boldsymbol{BGEBV}_t)}{sd(\boldsymbol{BGEBV}_t)}$, where $\boldsymbol{BGEBV}_t$ denotes the vector of GEBVs of base animals represented by cows with phenotypes.

Figure 3 shows the average differences in the number of progeny of the validation bulls between **P_2010** and **P_4020** scenarios relative to **P_Real**. The average difference in the number of progeny for **P_2010** was 4 (sires born in 2017), 2 (sires born in 2018), and 7 (sires born in 2019) for genotyped daughters and 1 (sires born in 2017), 2 (sires born in 2018), and 2 (sires born in 2019) for ungenotyped daughters, respectively. For **P_4020,** the differences were 10 (sires born in 2017), 5 (sires born in 2018), and 22 (sires born in 2019) for genotyped daughters and 3 (sires born in 2017), 4 (sires born in 2018), and 4 (sires born in 2019) for ungenotyped daughters. The highest mean difference in the number of sons was observed for 2017 for **P_4020**, genotyped bulls (1 son), whereas for ungenotyped bulls, for **P_2010** and **P_4020**, it was 0 for 2019.

Results

Validation

The validation model was computed for 562 validation bulls (387 genotyped and 175 ungenotyped) and 482,810 validation cows (30,227 genotyped and 452,336 ungenotyped). Figure 4A shows the estimated slopes (b1) of the validation models divided by scenario, sex, parity, and genotyping status. For bulls, the value of b_1 was close to the expected value of one for most scenarios, except **P_2010** (1.271) and **P_4020** (1.328) for **MF** for ungenotyped bulls. Moreover, for MF, overdispersion was observed when comparing the predictions of P_2010 and P_4020 with P_Real. Figure 4B shows intercepts (b_0) for all scenarios. Similar estimates, close to zero, were observed for every scenario, where the minimum value was -0.463 (**P_2010** for **RP** for genotyped bulls) and the maximum value was 0.067 (**P_4020** for **MF** for genotyped cows). In general, the intercepts estimated for bulls were negative, whereas for cows, the values were very close to zero. Figure 4C shows R^2,and Figure 4D displays the Pearson correlation coefficients between GEBVs predicted from the truncated and full

datasets for the validation animals, divided by scenario, sex, parity, and genotyping status. When **MF** or **GG** were used, the R^2 and correlations for cows were generally higher than those for the bulls. Additionally, the GEBV of genotyped cows always resulted in higher correlations than those of ungenotyped cows, regardless of missing parent handling. However, for **P_Real**, R^2 and correlations depended on the adopted missing parent approach. The most striking scenario was for the ungenotyped **P_Real**, where the correlation for **RP** was 0.892, decreased to 0.870 for **GG**, but increased to 0.924 for **MF**. The correlations for bulls followed a different pattern. For ungenotyped bulls, the correlation in each case increased sequentially from **RP** through **GG** to **MF**. For genotyped bulls, for **P_Real,** the correlation was similar across scenarios, but for **P_2010** and **P_4020, MF** resulted in slightly lower correlations.

Tables A1-A4 in the appendix provide detailed validation results for the genotyped and ungenotyped validation bulls and cows, respectively.

GEBV comparison

Figures 5 and 6 show GEBVs predicted for the full and the truncated datasets for dams and sires. For each scenario, the points on the scatter plots were grouped into elliptical clusters along the diagonal. However, for ungenotyped individuals, especially dams, we observed greater dispersion along the diagonal line. For ungenotyped bulls under the **P_4020** scenarios, GEBVs predicted from the truncated dataset were underestimated for bulls with the highest GEBSs in the full data, regardless of the applied approach of missing parent handling. Figure 7 shows the average GEBV trends for all scenarios divided by sex. The average GEBV trend is a combination of the population average (genetic trend: year-to-year change in the mean breeding value) and the intensity of selection (the advantage of selected parents within the year). After 2010, we observed that the mean GEBV increased for all animals, especially bulls. For each scenario, the highest increase in the GEBV trend was realized by **P_Real** with missing parents handled via **MF.**

Discussion

Our study focused on the comparison of different methods for expressing missing parental information (**RP**, **GG**, or **MF**) in the context of a single-step evaluation using the SNP-BLUP test-day model. In addition, we investigated the robustness of these three approaches to the increasing amount of missing parental information in the pedigree, separately for the genotyped and ungenotyped parts of the population. The analysis was conducted based on the real pedigree representing the Polish dairy cattle population, using fat yield as an example, and was assessed in terms of the statistical properties underlying the GEBV validation model. Two aspects emerge from these comparisons. First, the availability of genotype information enhances the quality of GEBVs prediction not only by resulting in "better" estimates of the parameters of the regression line in validation but also in higher R^2 and predicted versus observed GEBVs' correlations, as compared to ungenotyped individuals. Owing to the availability of genomic information, the choice of the method of missing parent handling does not impact GEBV prediction quality. The second aspect demonstrated in our study is that for ungenotyped individuals, the **MF** approach is not robust with respect to the number of missing parents. As the percentage of metafounders increases, it struggles to pass genotypic information to ungenotyped individuals. In particular, we observed that using metafounders led to the systematic underestimation of the GEBVs of the ungenotyped bulls with an increasing number of incomplete pedigree data, and consequently to an elevated b_1 in the validation model. Some, albeit not very strong, advantages of using **MF** over **GG** were previously reported, in the context of a single step G-BLUP evaluation, by Garcia-Baccino et al. (2017), Macedo et al. (2020), Macedo et al. (2022), and Kluska et al. (2021). The first study was based on simulated data and reported good predictive performance using **MF**. For the dairy sheep population, Macedo et al. (2020, 2022) obtained the most accurate GEBV prediction results, expressed by a validation slope close to unity, with **MF**. Kluska et al.

(2021) reported that in a beef cattle population, using **MF** and **GG** provided very similar results, but finally **MF** was ultimately recommended as the approach providing the smallest bias of GEBVs.

Conclusions

In summary, this study demonstrates that methods for handling missing parents in pedigrees may impact GEBV and that handling missing parents is increasingly important with the increasing number of incomplete pedigrees. The most important result of this study is that using the metafounder approach may lead to biased predictions for ungenotyped individuals, particularly as the proportion of missing parents increases. In contrast, for genotyped individuals, no marked differences in the handling of missing parent data were observed.

References

1. Bradford, H., Masuda, Y., VanRaden, P., Legarra, A., & Misztal, I. (2019). Modeling missing pedigree in single-step genomic BLUP. Journal of Dairy Science, 102(3), 2336–2346. https://doi.org/10.3168/jds.2018-15434
2. Garcia-Baccino, C. A., Legarra, A., Christensen, O. F., Misztal, I., Pocrnic, I., Vitezica, Z. G., & Cantet, R. J. C. (2017). Metafounders are related to F st fixation indices and reduce bias in single-step genomic evaluations. Genetics Selection Evolution, 49(1). https://doi.org/10.1186/s12711-017-0309-2
3. Himmelbauer, J., Schwarzenbacher, H., Fuerst, C., & Fuerst-Waltl, B. (2024). Exploring unknown parent groups and metafounders in single-step genomic best linear unbiased prediction: Insights from a simulated cattle population. Journal of Dairy Science, 107(10), 8170–8192. https://doi.org/10.3168/jds.2024-24891

4. Kluska, S., Masuda, Y., Ferraz, J. B. S., Tsuruta, S., Eler, J. P., Baldi, F., & Lourenco, D. (2021). Metafounders may reduce bias in composite cattle genomic prediction. Frontiers in Genetics, 12. https://doi.org/10.3389/fgene.2021.678587
5. Kudinov, A., Mäntysaari, E., Aamand, G., Uimari, P., & Strandén, I. (2020). Metafounder approach for single-step genomic evaluations of Red Dairy Cattle. Journal of Dairy Science, 103(7), 6299–6310. https://doi.org/10.3168/jds.2019-17483
6. Legarra, A., Bertrand, J., Strabel, T., Sapp, R., Sánchez, J., & Misztal, I. (2007). Multi-breed genetic evaluation in a Gelbvieh population. Journal of Animal Breeding and Genetics, 124(5), 286–295. https://doi.org/10.1111/j.1439-0388.2007.00671.x
7. Legarra, A., Christensen, O. F., Vitezica, Z. G., Aguilar, I., & Misztal, I. (2015). Ancestral relationships using metafounders: finite ancestral populations and across population relationships. Genetics, 200(2), 455–468. https://doi.org/10.1534/genetics.115.177014
8. Liu, Z., Reinhardt, F., Bünger, A., & Reents, R. (2004). Derivation and calculation of approximate reliabilities and Daughter Yield-Deviations of a random Regression Test-Day model for genetic evaluation of dairy cattle. Journal of Dairy Science, 87(6), 1896–1907. https://doi.org/10.3168/jds.s0022-0302(04)73348-2
9. Macedo, F., Astruc, J., Meuwissen, T., & Legarra, A. (2022). Removing data and using metafounders alleviates biases for all traits in Lacaune dairy sheep predictions. Journal of Dairy Science, 105(3), 2439–2452. https://doi.org/10.3168/jds.2021-20860
10. Macedo, F. L., Christensen, O. F., Astruc, J., Aguilar, I., Masuda, Y., & Legarra, A. (2020). Bias and accuracy of dairy sheep evaluations using BLUP and SSGBLUP with metafounders and unknown parent groups. Genetics Selection Evolution, 52(1). https://doi.org/10.1186/s12711-020-00567-1

11. Mäntysaari, E., Liu, Z., & VanRaden, P. (2010). Interbull validation test for genomic evaluations. Bulletin - International Bull Evaluation Service/Interbull Bulletin, 41, 17. https://journal.interbull.org/index.php/ib/article/download/1134/1125
12. Masuda, Y., Tsuruta, S., Bermann, M., Bradford, H. L., & Misztal, I. (2021). Comparison of models for missing pedigree in single-step genomic prediction. Journal of Animal Science, 99(2). https://doi.org/10.1093/jas/skab019
13. Strandén, Ismo. (2014). RelaX2 program for pedigree analysis, User's guide for version 1.65.
14. Westell, R., Quaas, R., & Van Vleck, L. (1988). Genetic groups in an animal model. Journal of Dairy Science, 71(5), 1310–1318. https://doi.org/10.3168/jds.s0022-0302(88)79688-5
15. Vandenplas, J., Veerkamp, R., Calus, M., Lidauer, M., Strandén, I., Taskinen, M., Schrauf, M., & Napel, J. T. (2022). 358. MiXBLUP 3.0 – software for large genomic evaluations in animal breeding programs. https://doi.org/10.3920/978-90-8686-940-4_358

Tables

Table 1 Number of animals in the analyzed dataset

Data	Sex	Number of animals	Number of records
Phenotype Full dataset (fat yield)	Cows	3,707,727	63,615,019
Phenotype Truncated dataset (fat yield)		3,224,917	58,446,695
Genotype	Cows	113,019	181,991
	Bulls	68,972	
Pedigree	Cows	4,569,044	4,712,143
	Bulls	143,099	

Figures

Figure 1 Percentage of missing parents for each pedigree scenario.

Figure 2 Number of individuals in each genetic group for each pedigree scenario.

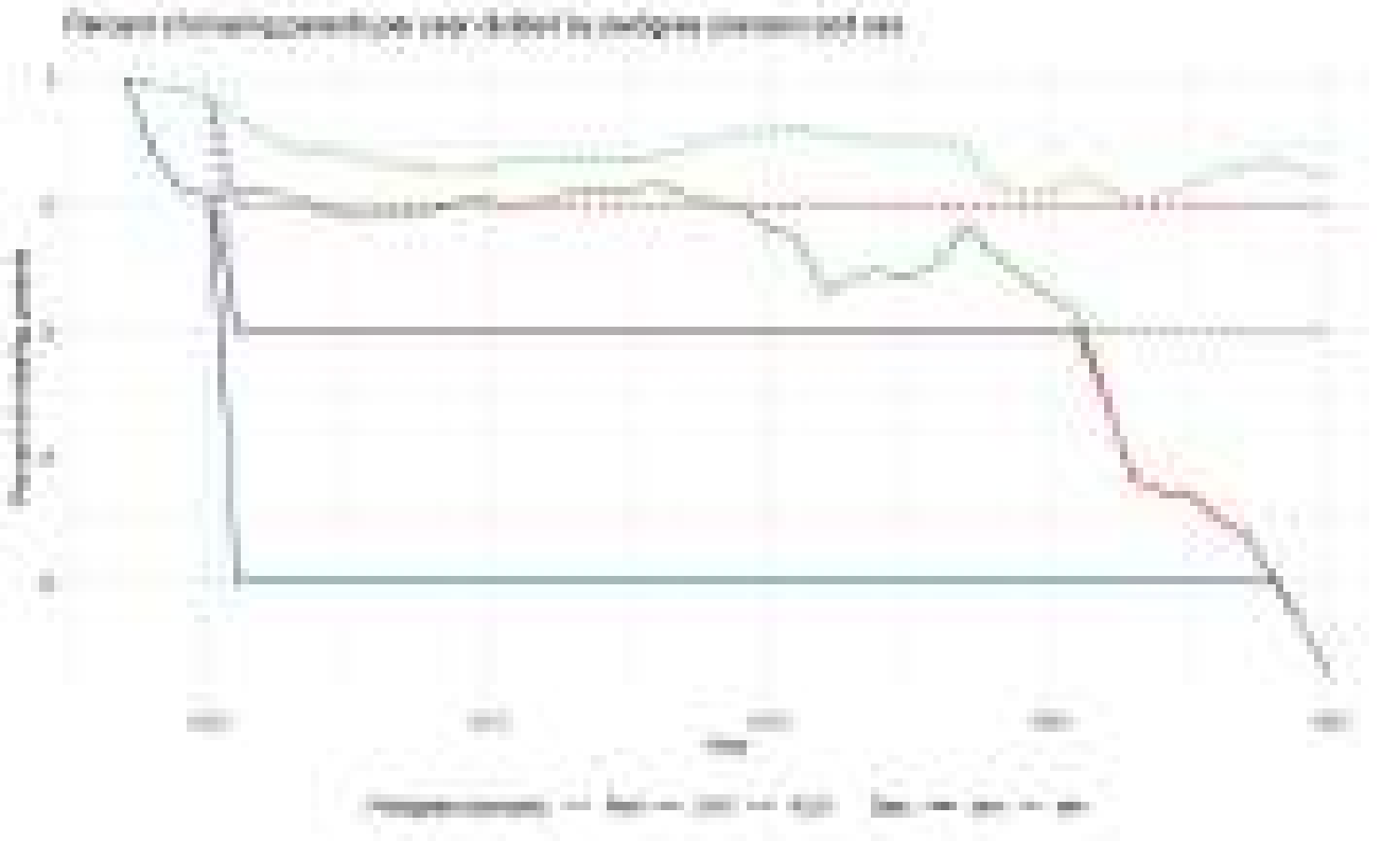
Figure 3 Average difference in the number of progenies between RP and 2010, 4020 divided by genotyped and ungenotyped individuals.

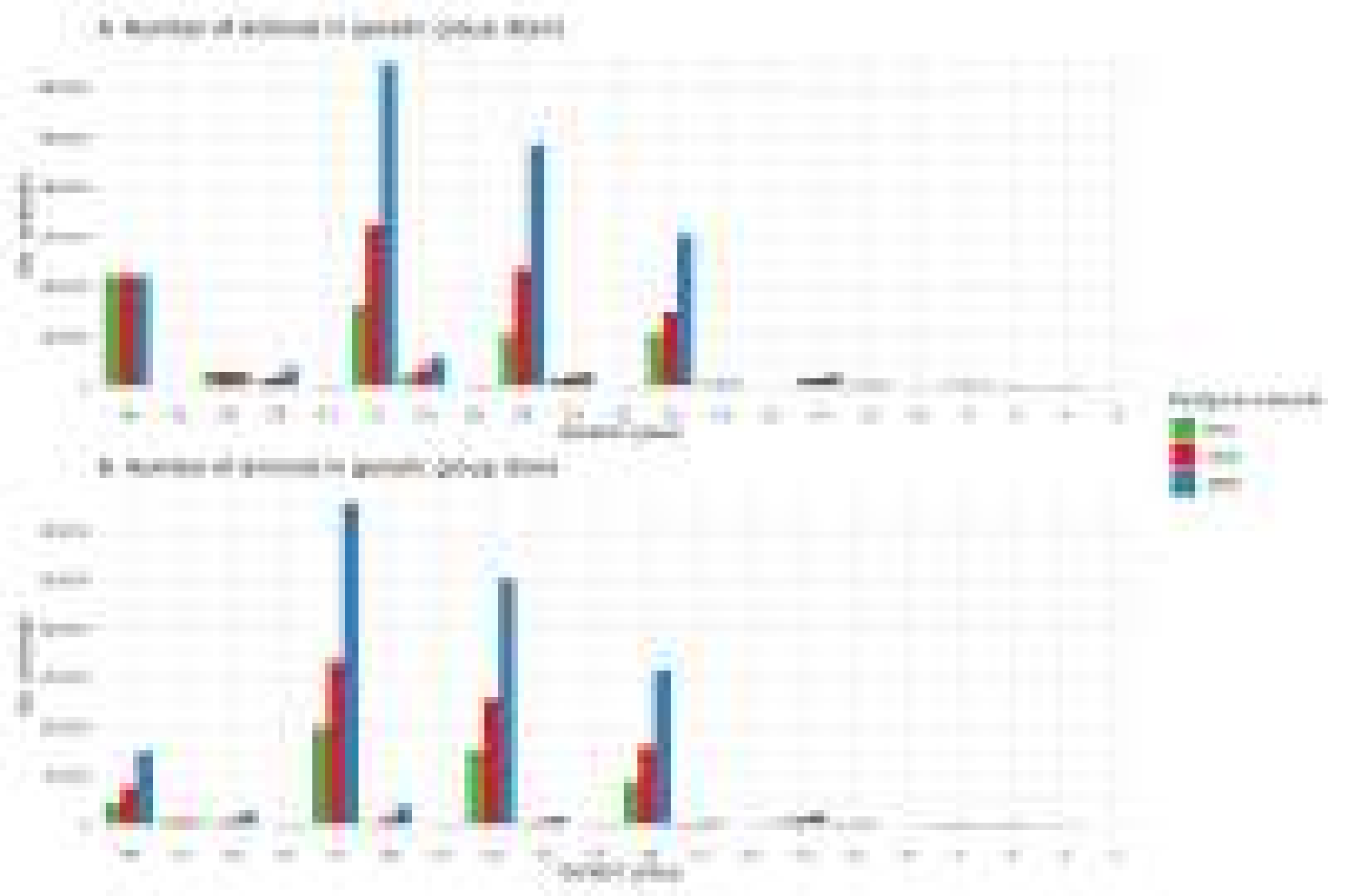
Figure 4 Validation results for individuals divided by sex and method.

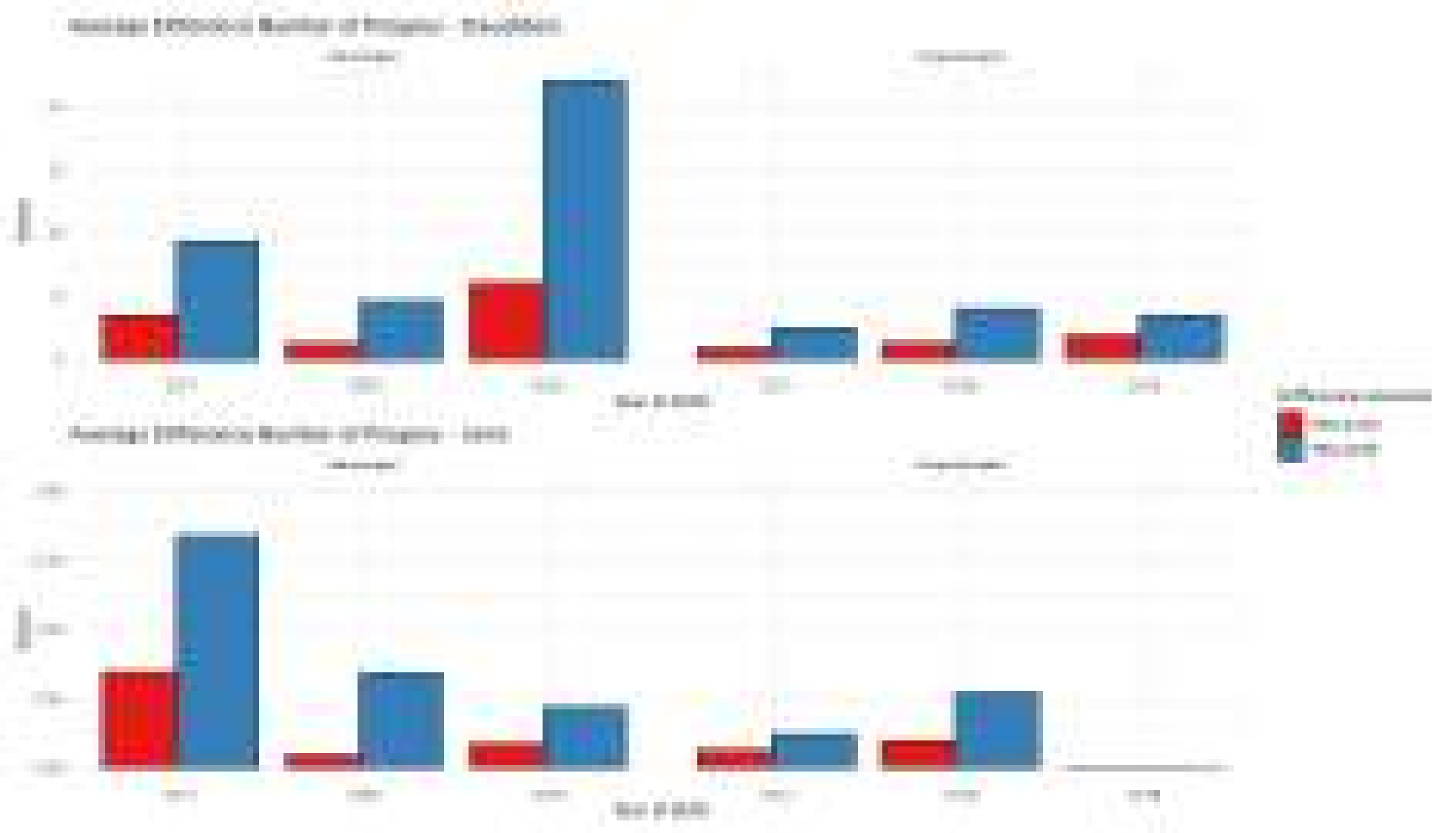
Figure 5 Comparison of GEBV for dams across scenarios divided into genotyped and ungenotyped individuals.

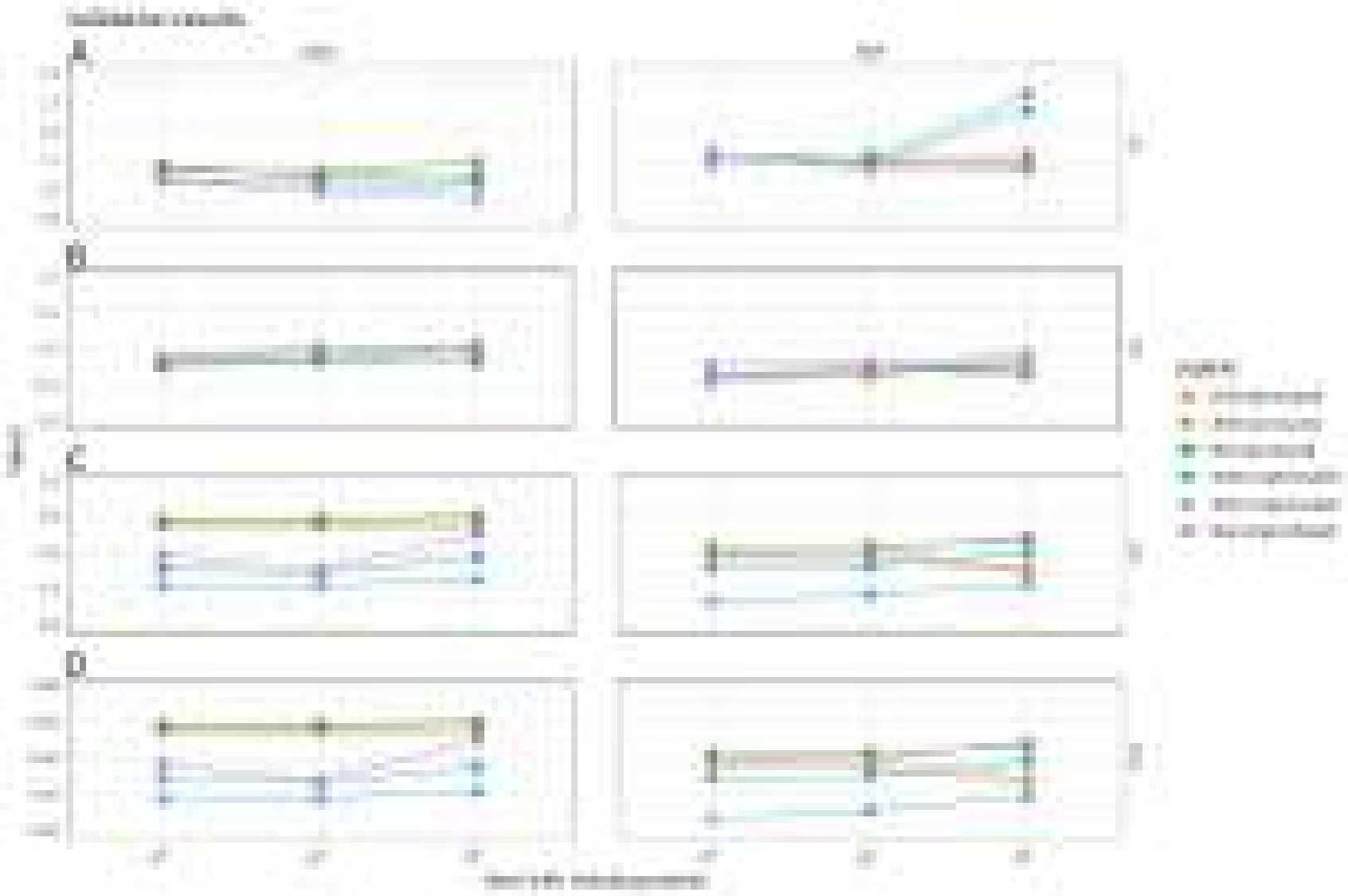
Figure 6 Comparison of GEBV for sires across scenarios divided into genotyped and ungenotyped individuals.

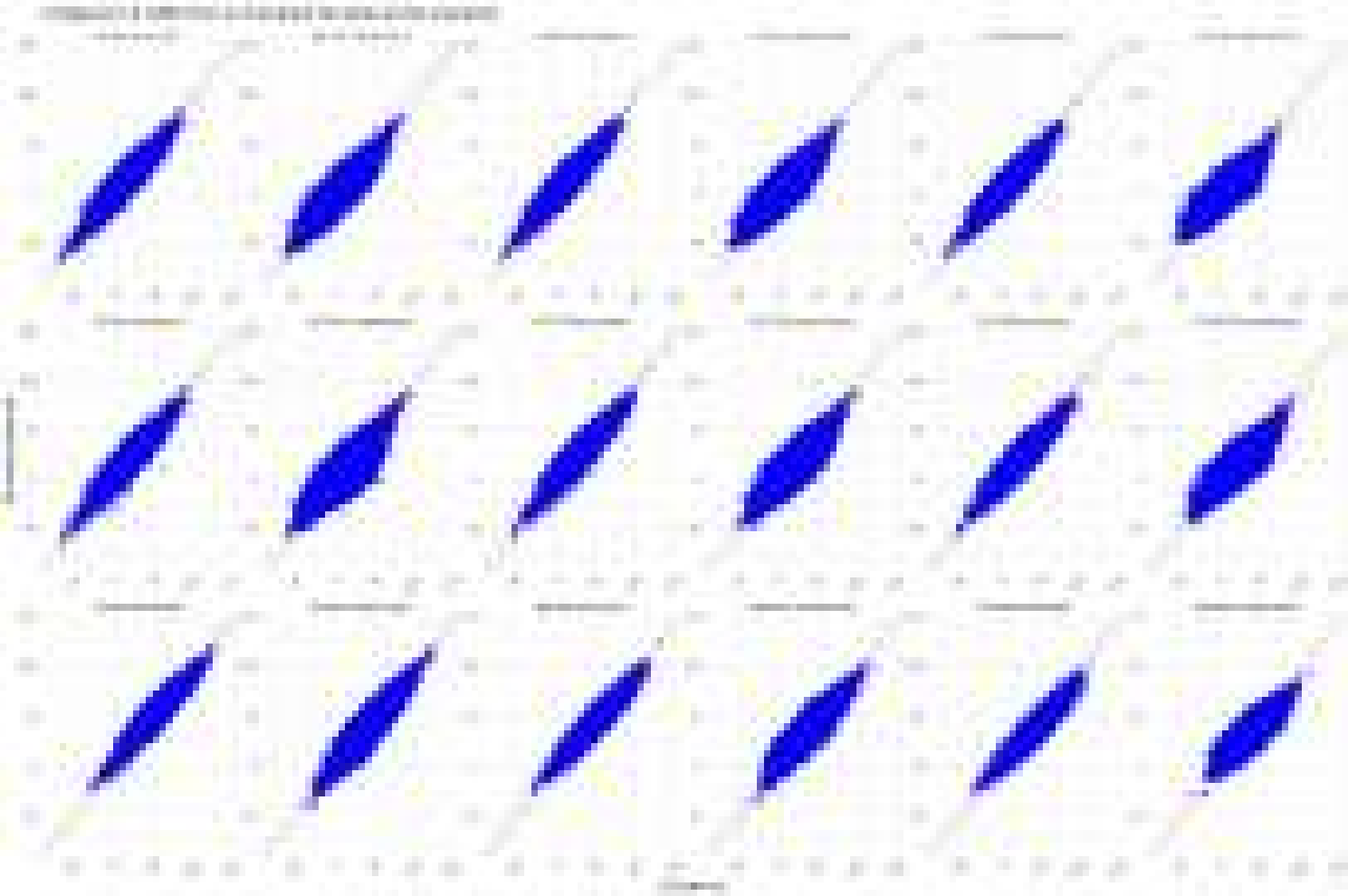
Figure 7 Genetic trends for individuals divided by sex and method.

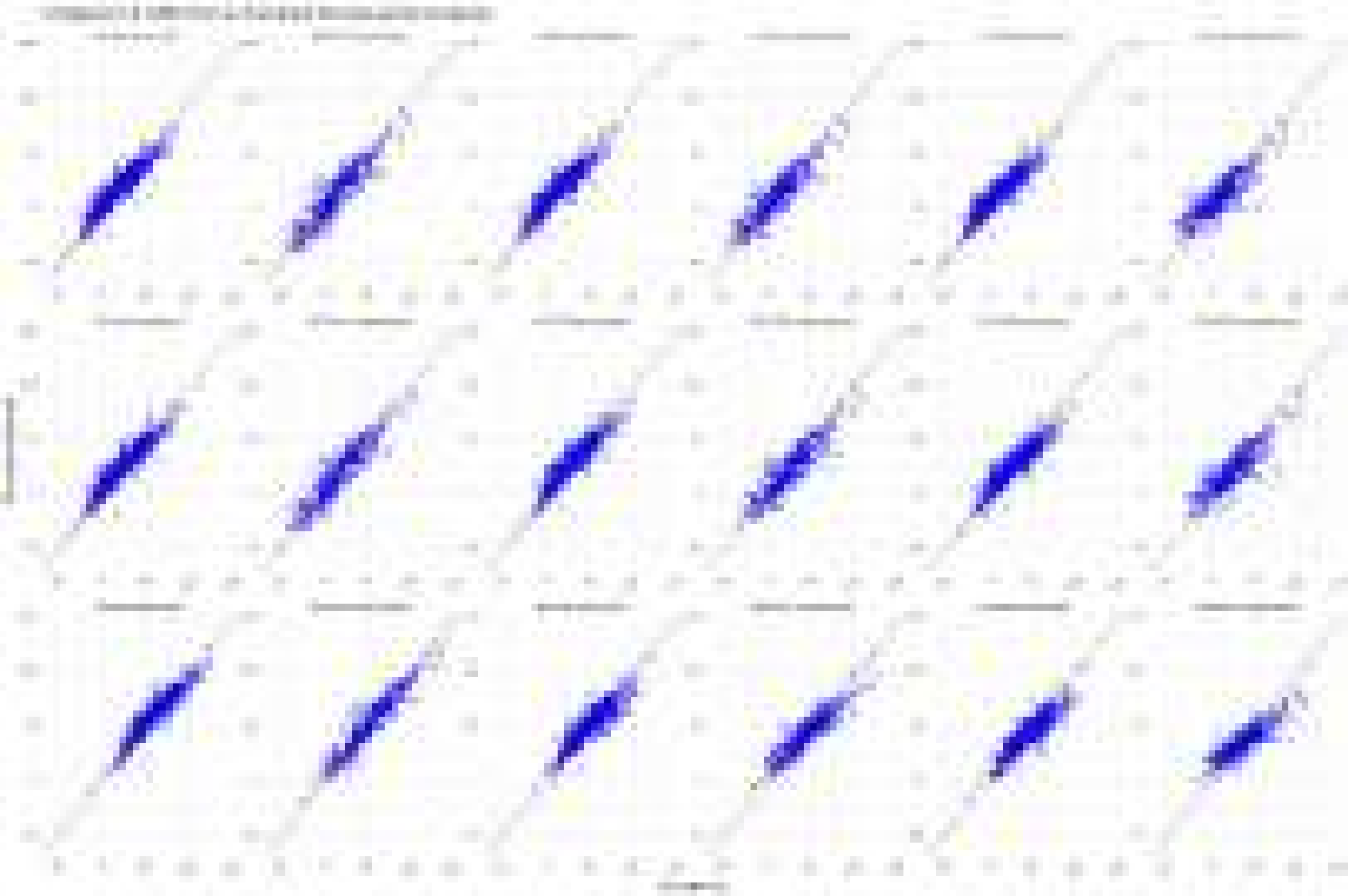


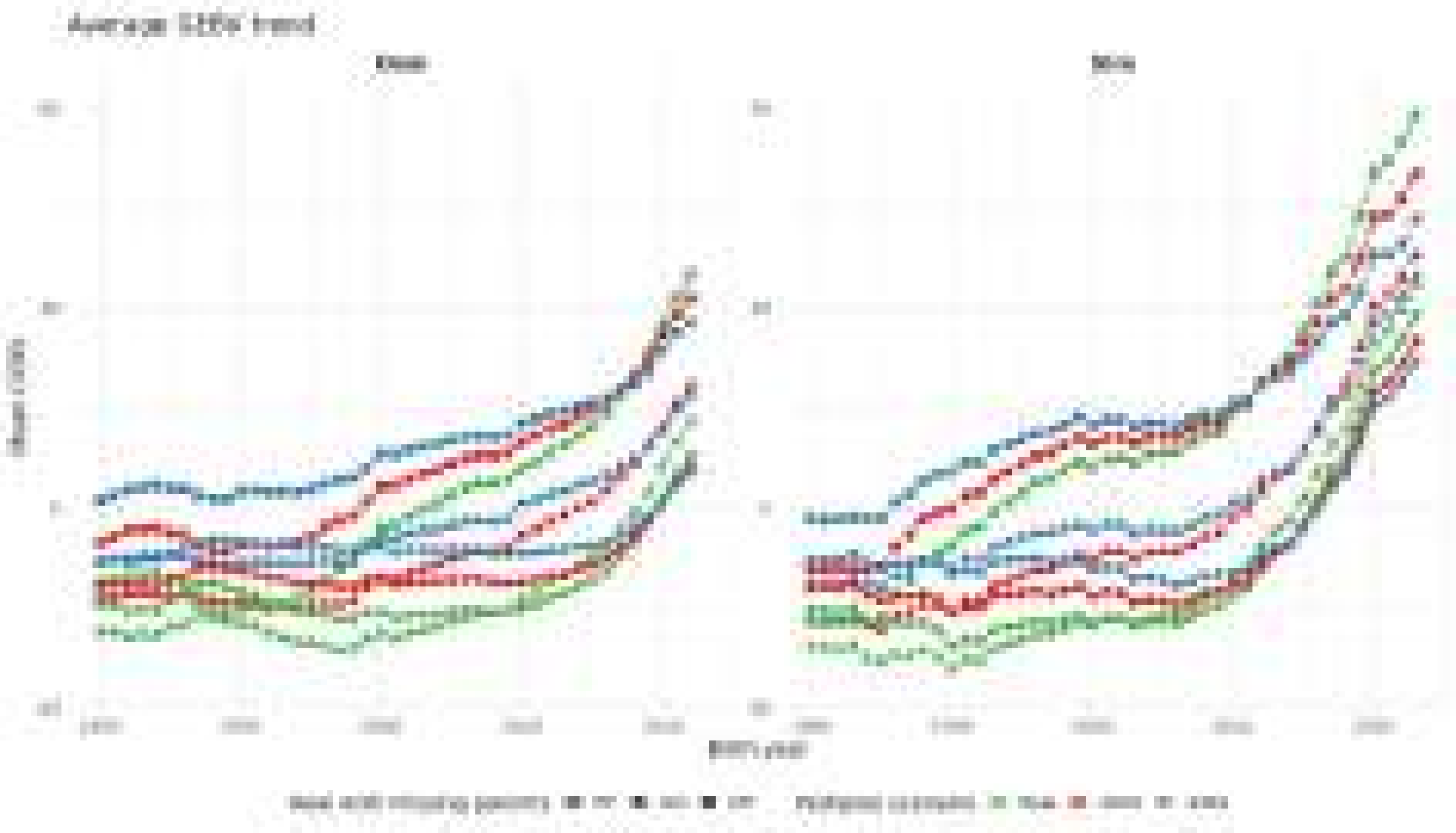





A
B
C
D






Average EEW trend
Male
Skin

OŚWIADCZENIE WSPÓŁAUTORA

Balice, 05.11.2025

Imię i nazwisko: **dr inż. Kacper Żukowski**

Afiliacja: **Zakład Hodowli Bydła**

Instytut Zootechniki Państwowy Instytut Badawczy w Krakowie

Oświadczenie

Oświadczam, że w pracy Słomian, D., Żukowski, K., Szyda, J. (2023). Heterogeneity in convergence behaviour of the single-step SNP-BLUP model across different effects and animal groups. Genetics Selection Evolution, 55(1). https://doi.org/10.1186/s12711-023-00856-5

mój udział polegał na: przygotowaniu danych genomowych, współtworzeniu, korekcie i edycji manuskryptu.

.....

(czytelny podpis współautora)

OŚWIADCZENIE WSPÓŁAUTORA

Wrocław, 05.12.2025

Imię i nazwisko: **Prof. dr hab. Joanna Szyda**

Afiliacja: **Wydział Biologii i Hodowli Zwierząt**

Uniwersytet Przyrodniczy we Wrocławiu

Oświadczenie

Oświadczam, że w pracy Słomian, D., Żukowski, K., Szyda, J. (2023). Heterogeneity in convergence behaviour of the single-step SNP-BLUP model across different effects and animal groups. Genetics Selection Evolution, 55(1). https://doi.org/10.1186/s12711-023-00856-5

mój udział polegał na: opracowywaniu metodyki, administracja projektem, nadzorem nad danymi i ich analizą, współtworzeniu, korekcie i edycji manuskryptu.

J. Szyda ……..

(czytelny podpis współautora)

OŚWIADCZENIE WSPÓŁAUTORA

Balice, 05.11.2025

Imię i nazwisko: **dr inż. Kacper Żukowski**

Afiliacja: **Zakład Hodowli Bydła**

Instytut Zootechniki Państwowy Instytut Badawczy w Krakowie

Oświadczenie

Oświadczam, że w pracy Słomian, D., Żukowski, K., Szyda, J. (2025). A comparison of genomically enhanced breeding values predicted by different single-step approaches. Annals of Animal Science. https://doi.org/10.2478/aoas-2025-0088

mój udział polegał na: przygotowaniu danych genomowych, współtworzeniu, korekcie i edycji manuskryptu.

.....Żuk Kap.....

(czytelny podpis współautora)

OŚWIADCZENIE WSPÓŁAUTORA

Wrocław, 05.12.2025

Imię i nazwisko: **Prof. dr hab. Joanna Szyda**

Afiliacja: **Wydział Biologii i Hodowli Zwierząt**

Uniwersytet Przyrodniczy we Wrocławiu

Oświadczenie

Oświadczam, że w pracy Słomian, D., Żukowski, K., Szyda, J. (2025). A comparison of genomically enhanced breeding values predicted by different single-step approaches. Annals of Animal Science. https://doi.org/10.2478/aoas-2025-0088

mój udział polegał na: opracowywaniu metodyki, administracja projektem, nadzorem nad danymi i ich analizą, współtworzeniu, korekcie i edycji manuskryptu.

J. Szyda

……..

(czytelny podpis współautora)

OŚWIADCZENIE WSPÓŁAUTORA

Wrocław, 05.12.2025

Imię i nazwisko: **mgr inż. Michalina Jakimowicz**

Afiliacja: **Wydział Biologii i Hodowli Zwierząt**

Uniwersytet Przyrodniczy we Wrocławiu

Oświadczenie

Oświadczam, że w pracy Słomian, D., Jakimowicz, M., Suchocki, T., Szyda, J. (2025). Comparison of BLUPF90IOD3 and MiXBLUP implementations of the single-step model applied to the Polish national dairy cattle evaluation. PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-8398690/v1]

mój udział polegał na: edycji danych, przeprowadzeniu analiz, współtworzeniu, korekcie i edycji manuskryptu.

Michalina Jakimowicz

(czytelny podpis współautora)

OŚWIADCZENIE WSPÓŁAUTORA

Wrocław, 05.12.2025

Imię i nazwisko: **dr hab. inż. Tomasz Suchocki**

Afiliacja: **Wydział Biologii i Hodowli Zwierząt**

Uniwersytet Przyrodniczy we Wrocławiu

Oświadczenie

Oświadczam, że w pracy Słomian, D., Jakimowicz, M., Suchocki, T., Szyda, J. (2025). Comparison of BLUPF90IOD3 and MiXBLUP implementations of the single-step model applied to the Polish national dairy cattle evaluation. PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-8398690/v1]

mój udział polegał na: przeprowadzeniu analiz, współtworzeniu, korekcie i edycji manuskryptu.

T. Suchocki

(czytelny podpis współautora)

OŚWIADCZENIE WSPÓŁAUTORA

Wrocław, 05.12.2025

Imię i nazwisko: **Prof. dr hab. Joanna Szyda**

Afiliacja: **Wydział Biologii i Hodowli Zwierząt**

Uniwersytet Przyrodniczy we Wrocławiu

Oświadczenie

Oświadczam, że w pracy Słomian, D., Jakimowicz, M., Suchocki, T., Szyda, J. (2025). Comparison of BLUPF90IOD3 and MiXBLUP implementations of the single-step model applied to the Polish national dairy cattle evaluation. PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-8398690/v1]

mój udział polegał na: opracowywaniu metodyki, administracja projektem, nadzorem nad danymi i ich analizą, współtworzeniu, korekcie i edycji manuskryptu.

J. Szyda

……..

(czytelny podpis współautora)

OŚWIADCZENIE WSPÓŁAUTORA

Wageningen, 05.12.2025

Imię i nazwisko: **dr inż. Jeremie Vandenplas**

Afiliacja: **Animal Breeding & Genomics**

Wageningen University & Research

Oświadczenie

Oświadczam, że w pracy Słomian, D., Vandenplas, J., Ten Napel, J., Żukowski, K., Skarwecka, M., Szyda, J. (2025). Modeling missing parents in single-step test-day SNP-BLUP evaluation of dairy cattle. [Preprint]. bioRxiv. https://doi.org/10.64898/2025.12.02.691779

mój udział polegał na: opracowywaniu metodyki, nadzorem nad analizą danych, współtworzeniu, korekcie i edycji manuskryptu.

.....

(czytelny podpis współautora)

OŚWIADCZENIE WSPÓŁAUTORA

Wageningen, 05.12.2025

Imię i nazwisko: **dr inż. Jan Ten Napel**

Afiliacja: **Animal Breeding & Genomics**

Wageningen University & Research

Oświadczenie

Oświadczam, że w pracy Słomian, D., Vandenplas, J., Ten Napel, J., Żuławski, K., Skarwecka, M., Szyda, J. (2025) Modeling missing parents in single-step test-day SNP-BLUP evaluation of dairy cattle.(Preprint) bioRxiv. https://doi.org/10.64898/2025.12.02.691779

mój udział polegał na: opracowywaniu metodyki, nadzorem nad analizą danych, współtworzeniu, korekcie i edycji manuskryptu.

.....

(czytelny podpis współautora)

OŚWIADCZENIE WSPÓŁAUTORA

Balice, 05.11.2025

Imię i nazwisko: **dr inż. Kacper Żukowski**

Afiliacja: **Zakład Hodowli Bydła**

Instytut Zootechniki Państwowy Instytut Badawczy w Krakowie

Oświadczenie

Oświadczam, że w pracy Słomian, D., Vandenplas, J., Ten Napel, J., Żukowski, K., Skarwecka, M., Szyda, J. (2025). Modeling missing parents in single-step test-day SNP-BLUP evaluation of dairy cattle. [Preprint]. bioRxiv. https://doi.org/10.64898/2025.12.02.691779

mój udział polegał na: przygotowaniu danych genomowych, współtworzeniu, korekcie i edycji manuskryptu.

.....

(czytelny podpis współautora)

OŚWIADCZENIE WSPÓŁAUTORA

Balice, 05.12.2025

Imię i nazwisko: **dr inż. Monika Skarwecka**

Afiliacja: **Zakład Hodowli Bydła**

Instytut Zootechniki Państwowy Instytut Badawczy w Krakowie

Oświadczenie

Oświadczam, że w pracy Słomian, D., Vandenplas, J., Ten Napel, J., Żukowski, K., Skarwecka, M., Szyda, J. (2025). Modeling missing parents in single-step test-day SNP-BLUP evaluation of dairy cattle. [Preprint]. bioRxiv. https://doi.org/10.64898/2025.12.02.691779

mój udział polegał na: współtworzeniu, korekcie i edycji manuskryptu.

Monika Skarwecka

(czytelny podpis współautora)

OŚWIADCZENIE WSPÓŁAUTORA

Wrocław, 05.12.2025

Imię i nazwisko: **Prof. dr hab. Joanna Szyda**

Afiliacja: **Wydział Biologii i Hodowli Zwierząt**

Uniwersytet Przyrodniczy we Wrocławiu

Oświadczenie

Oświadczam, że w pracy Słomian, D., Vandenplas, J., Ten Napel, J., Żukowski, K., Skarwecka, M., Szyda, J. (2025). Modeling missing parents in single-step test-day SNP-BLUP evaluation of dairy cattle. [Preprint]. bioRxiv. https://doi.org/10.64898/2025.12.02.691779

mój udział polegał na: opracowywaniu metodyki, administracja projektem, nadzorem nad danymi i ich analizą, współtworzeniu, korekcie i edycji manuskryptu.

J. Szyda

……..

(czytelny podpis współautora)