Analiza statistică preliminară a unei colecții de date

efectuat în cadrul proiectului *Abordarea bioeconomică a agenților antimicrobieni – utilizare și rezistență*

(cod - PN-III-P1-1.2-PCCDI-2017-0361).

Colectiv de redacţie:

Raluca Mureşan: pre-procesare date, vizualizare şi sumarizare (sect. 2.1, sect. 2.2, sect. 2.3)

Claudia Zaharia: pre-procesare date, analiza de corelaţie (sect. 2.4, sect. 2.5)

Kristian Miok: agregarea datelor, analiza preliminară a pachetelor R folosite în studiile de asociere (sect. 3.1, sect 3.2, Anexa 1)

Daniela Zaharie: stabilirea structurii, redactare (sect. 1), validare finală

Coordonator: Daniela Zaharie

Membri: Raluca Mureşan, Claudia Zaharia, Kristian Miok

Data finalizării: 29.11.2018

Versiunea: 1.0

Acknowledgements

Activities under this work were carried out in the *Research Laboratory Complex "Horia Cernescu"* - financed by project *"A bio-economical approach of the antimicrobial agents - use and resistance"*, in the frame of contract PCCDI 7/19.03.2018, code: PN-III P1-1.2-FPRD-2017.

1. Colecţia de date

Scopul acestui raport este de a prezenta rezultate preliminare obţinute în urma analizei unei colecţii de date. Colecţia a fost analizată din perspectiva particularităţilor datelor şi a identificării tehnicilor statistice adecvate.

Colecţia de date conţine informaţii referitoare la producţia de lapte înregistrată în perioada 2012-2018 la Staţiunea de Cercetare-Dezvoltare pentru Creşterea Bovinelor Arad şi a fost pusă la dispoziţia echipei proiectului BioAMR cu scopul de a efectua analize preliminare.

1.1. Structura datelor

Colecţia de date este structurată pe ani iar pentru fiecare an este organizată în mai multe fişiere corespunzătoare următoarelor mărimi înregistrate: cantitate de lapte, compoziţie (lactoză, grăsime, cazeină, proteine, Ph, uree, raport grăsime-proteine, număr celule somatice, substanţă uscată negrasă - SUN). Pentru fiecare an există valori asociate cu maxim 13 măsurători. Numărul de animale pentru care s-au făcut înregistrări variază de la un an la altul după cum urmează:

An	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**
Nr măsurători/ an	10	12	12	12	12	13
Nr animale	368	345	380	404	386	383

Intrucât datele corespunzătoare anului 2016 conţin doar valori corespunzătoare pentru cantitatea de lapte, grăsime şi cazeină, acest an a fost exclus din analiză. Prin urmare au fost analizate date pentru anii 2012, 2013, 2014, 2015 şi 2017 (fişierele date 2012.csv – date2017.csv) scopul principal fiind acela de a analiza eventualele corelaţii dintre numărul de celule somatice (interpretat ca indicator al prezenţei unei infecţii) şi celelalte caracteristici.

1.2. Preprocesarea datelor

In vederea analizei de corelaţie, fişierele corespunzătoare fiecărui an au fost agregate într-un singur fişier având următoarele câmpuri: cod identificare animal, număr de ordine măsurătoare, cele 10 valori măsurate (cantitate, cazeină, grăsime, lactoză, număr celule somatice, ph, proteine, raport grăsime proteine, SUN – substanţa uscată negrasă, uree). Anexa 1 conţine un fragment din codul implementat în R folosit pentru extragerea datelor din fişierele iniţiale şi construirea colecţiilor de date utilizate în analiza exploratorie. In această etapă analiza a fost realizată separat pentru fiecare an.

În seturile de date au fost identificate un număr de câmpuri de valoare 0 (e.g., pentru variabilele *Ph, cantitate, grăsime, lactoză*), codificând date lipsă. Aceste date au fost eliminate în etapa de preprocesare.

2. Analiza exploratorie

2.1 Selecţia datelor incluse în analiză

Scopul analizei exploratorii a fost acela de a studia distribuţia parametrilor laptelui recoltat şi evoluţia acestora în timp. De un interes deosebit a fost numărul de celule somatice (*NCS*), întrucât acesta este un indicator important al calităţii laptelui ce reflectă prezenţa sau absenţa unei infecţii a ugerului. Au fost urmărite corelaţiile dintre *NCS* şi ceilalţi parametri de cantitate şi calitate a laptelui recoltat. Pentru a putea studia corelaţiile intra-individuale, din datele colectate pe fiecare an au fost incluse în analiză doar cele corespunzătoare vacilor pentru care au fost disponibile cel puţin 10 măsurători repetate.

2.2 Modalităţi de vizualizare a datelor

Principalele modalităţi de vizualizare a datelor cantitative sunt histograma şi diagrama boxplot.

Histograma arată distribuţia de frecvenţe a datelor, permiţând studiul vizual al simetriei, normalităţii, şi selecţia procedeelor de analiză statistică inferenţială ulterioară adecvate. Pentru exemplificare, Figura 1 reprezintă histograma variabilei *gras* pe anul 2012, care indică o distribuţie simetrică, aproximativ normală, cele mai frecvente valori fiind în intervalul [3.5, 4.5]. Figura 2, histograma pentru numărul de celule somatice din lapte pentru anul 2017, arată o distribuţie non-normală, cu pronunţată asimetrie stânga, majoritatea valorilor fiind foarte mici (indicând caracterul "excepţional" al cazurilor de infecţie).

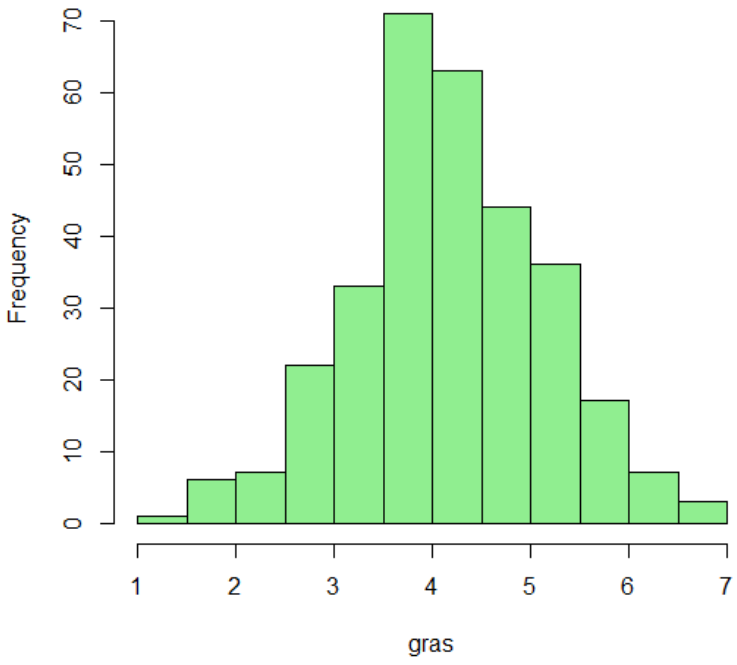


Figura 1: Histograma *gras* pe anul 2012

Contract 7PCCDI, COD: PN-III-P1-1.2-PCCDI-2017-0361

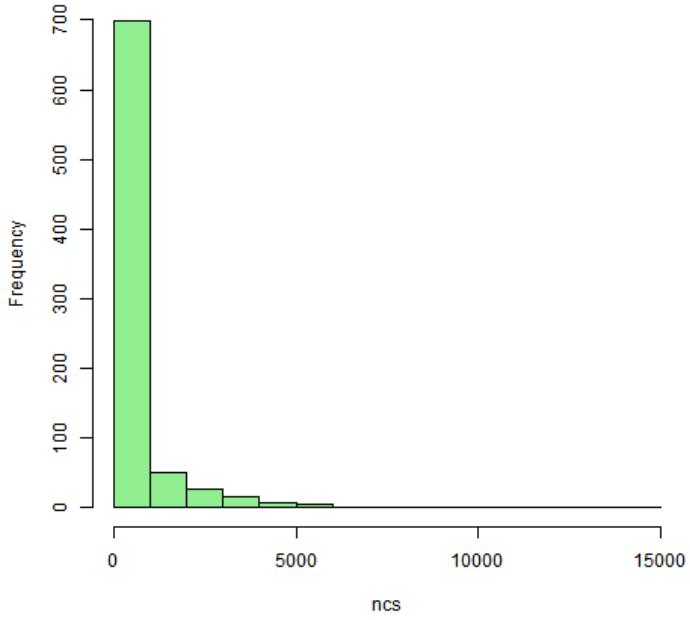


Figura 2: Histograma *NCS* pe anul 2017

Diagrama boxplot figurează grafic valori de sumar importante ale seriei de date (minim, maxim, cuartile, mediană) şi pune în evidenţă eventuale valori atipice, lucru care poate fi util în etapa de analiză preliminară pentru identificarea erorilor, sau ulterior pentru identificarea unor situaţii speciale. Este considerată valoare atipică orice valoare în afara intervalului $[Q1 - 1.5 \cdot IQR, Q3 + 1.5 \cdot IQR]$, unde $Q1$ şi $Q3$ reprezintă cuartila inferioară, respectiv superioară a seriei de date, iar $IQR = Q3 - Q1$ este lungimea intervalului intercuartilic (a se vedea Figura 3). De exemplu, în Figura 4 (boxplot pentru pH grupat pe sesiuni de măsurători, pe anul 2014) se poate observa prezenţa unei valori a pH-ului de 13.07, care constituie o eroare de măsurare (în consecinţă, valoarea a fost eliminată în analiza finală). Pe de altă parte, în Figura 5, valorile atipice (foarte mari) pentru numărul de celule somatice indică o probabilă infecţie a ugerului.

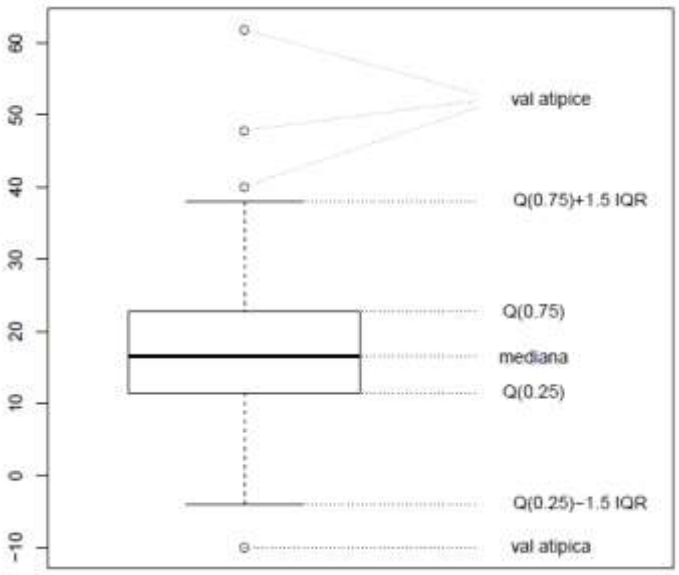


Figura 3: Elementele unui boxplot

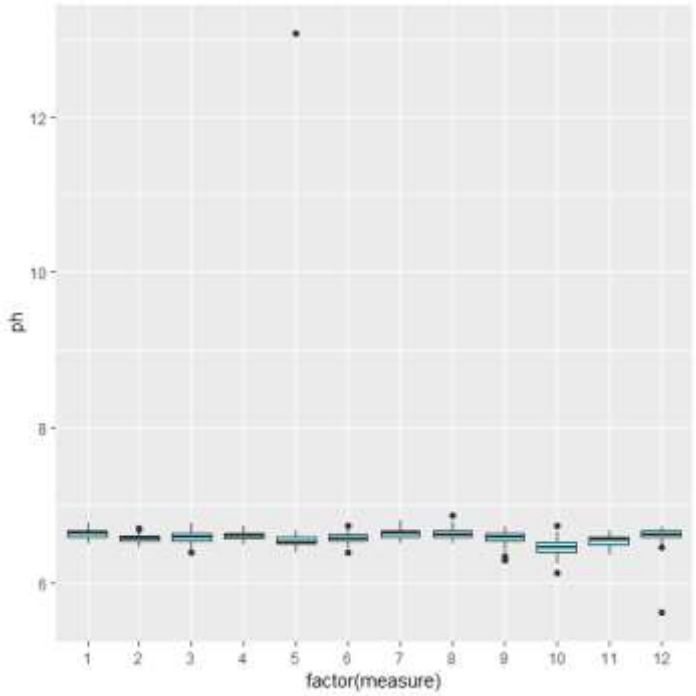


Figura 4: Diagrama boxplot pentru *ph* grupat pe sesiuni de măsurători pentru anul 2014 (date preliminare)

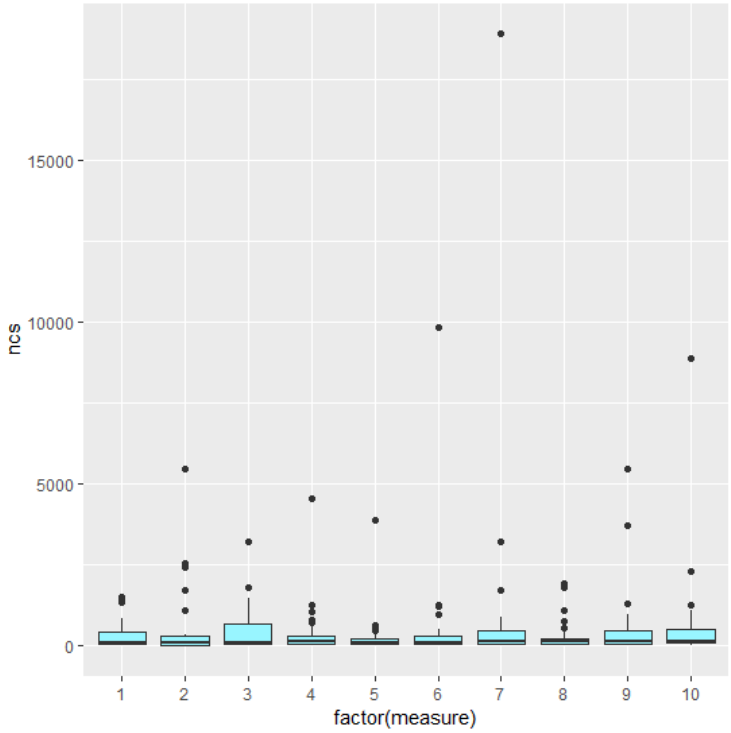


Figura 5: Diagrama boxplot pentru *ncs* grupat pe sesiuni de măsurători pentru anul 2012

Reprezentarea boxplot-urilor grupate pe sesiuni de măsurători permite de asemenea vizualizarea unor posibile tendinţe în evoluţia parametrilor între sesiuni. Spre exemplu, în Figura 6 se poate observa faptul că grăsimea din lapte este mai scăzută pe perioada verii şi mai ridicată pe perioada iernii.

Contract 7PCCDI, COD: PN-III-P1-1.2-PCCDI-2017-0361

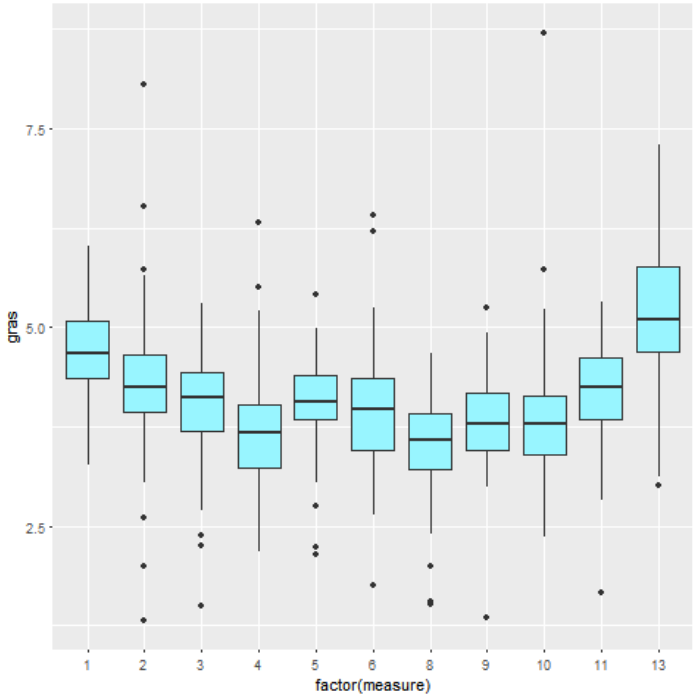


Figura 6: Diagrama boxplot pentru mărimea *gras* grupată pe sesiuni de măsurători pentru anul 2017

Histograme şi diagrame boxplot pentru toţi parametrii analizaţi şi pentru toţi anii de studiu pot fi găsite în arhiva AnalizaExploratorie.rar.

2.3 Modalităţi de sumarizare a datelor

Pentru sumarizarea datelor cantitative prin valori numerice se pot utiliza indici de poziţie sau grupare (*medie*, *mediană*, *p-cuantile*, *cuartile*), respectiv indici de împrăştiere (*deviaţie standard*, *lungimea intervalului intercuartilic*). Dintre acestea, cele mai utilizate sunt media şi deviaţia standard.

Pentru un eşantion de date $(x_1, x_2, \dots, x_n)$, aceste mărimi sunt definite după cum urmează:

- Media de eşantion este $\bar{x} = \frac{x_1+x_2+\cdots+x_n}{n}$
- Mediana este valoarea aflată la mijlocul seriei ordonate de date (mai mare decât ½ din date şi mai mică decât celelalte ½ din date)
- p-cuantila ($p \in (0,1)$) este valoarea mai mare decât o proporţie p din date şi mai mică decât o proporţie 1-p din date (de ex. mediana este 0.5 – cuantila)
- Cuartilele sunr p-cuantile particulare, obţinute pentru p=0.25 (cuartila inferioară, notată $Q1$), respectiv p=0.75 (cuartila superioară, notată $Q3$)
- Deviaţia standard de eşantion se defineşte prin $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- Lungimea intervalului intercuartilic $IQR = Q3 - Q1$

Valori mari pentru s, respectiv IQR, indică un grad mare de variabilitate a datelor, în timp ce valori apropiate de 0 indică date omogene.

În Tabelul 1 sunt prezentate valori numerice de sumar pentru fiecare dintre mărimile corespunzătoare setului de date folosind valori înregistrate corespunzătoare celor 5 ani de studiu, precum şi p-valorile testelor Shapiro-Wilk de normalitate (o valoare a lui *p* mai mică decât 0.05 indică o distribuţie non-normală a datelor).

Se observă că mărimile *gras, lact, ph, prot, gras_prot* şi *sun* nu prezintă fluctuaţii semnificative de la un an la altul, mediile şi deviaţiile standard rămânând aproximativ constante. Singura observaţie care rezultă din analiza indicatorilor statistici calculaţi se referă la cantitatea medie de lapte recoltată care este mai scăzută în 2014 faţă de ceilalţi ani.

În ceea ce priveşte parametrul de interes *ncs*, acesta prezintă o mare variabilitate, reflectată într-o plajă largă de valori, deviaţie standard mare, precum şi fluctuaţii semnificative de la un an la altul.

Se constată de asemenea că distribuţiile a doar trei din seriile de date analizate sunt normale (*gras* pentru 2012 şi *sun* pentru 2012 şi 2014), situaţia predominantă fiind de non-normalitate. În cazul în care se doreşte efectuarea de comparaţii statistice pe variabile între ani, acest fapt recomandă utilizarea de *tehnici neparametrice de analiză*.

Variabila	*An*	*Min*	*Max*	$\bar{x}$	*s*	*p*
cant	2012	2.00	17.20	8.18	2.56	0.009
	2013	1.50	17.00	8.29	2.70	<0.001
	2014	1.50	15.50	6.95	2.02	<0.001
	2015	1.50	16.00	8.16	2.39	0.001
	2017	1.80	15.20	8.49	2.41	0.034
gras	2012	1.31	6.94	4.17	0.99	0.632
	2013	1.47	9.21	4.16	0.98	<0.001
	2014	1.29	7.52	4.27	0.82	<0.001
	2015	0.71	8.93	4.06	0.91	<0.001
	2017	1.32	8.70	4.09	0.83	<0.001
lact	2012	3.29	5.12	4.63	0.28	<0.001
	2013	3.28	5.16	4.69	0.21	<0.001
	2014	3.16	5.26	4.67	0.25	<0.001
	2015	2.82	5.37	4.77	0.23	<0.001
	2017	3.34	5.26	4.79	0.19	<0.001
ncs	2012	6	18907	514.80	1471.96	<0.001
	2013	2	14787	431.58	1188.48	<0.001
	2014	2	13797	419.39	1075.44	<0.001
	2015	0	13235	290.20	773.90	<0.001
	2017	8	14988	523.68	1074.38	<0.001
ph	2012	5.93	6.73	6.52	0.08	<0.001
	2013	6.22	6.83	6.62	0.07	<0.001
	2014	5.61	6.86	6.57	0.09	<0.001
	2015	5.93	6.78	6.58	0.08	<0.001
	2017	6.03	6.83	6.57	0.07	<0.001
prot	2012	2.25	5.39	3.63	0.49	0.009
	2013	2.43	5.69	3.56	0.51	<0.001
	2014	2.25	5.32	3.62	0.46	<0.001
	2015	2.06	6.04	3.52	0.53	<0.001
	2017	2.33	6.40	3.58	0.51	<0.001

gras_prot	2012	0.32	1.68	1.15	0.22	<0.001
	2013	0.43	2.38	1.17	0.26	<0.001
	2014	0.33	2.71	1.18	0.21	<0.001
	2015	0.24	2.32	1.16	0.23	<0.001
	2017	0.37	2.38	1.15	0.21	<0.001
sun	2012	7.39	10.47	9.10	0.49	0.79
	2013	7.64	11.42	8.97	0.56	<0.001
	2014	7.79	11.01	9.09	0.46	0.169
	2015	7.42	10.81	9.14	0.50	<0.001
	2017	7.92	11.21	9.17	0.44	<0.001
uree	2012	0.40	70.00	26.97	12.74	0.028
	2013	1.30	64.10	19.30	10.33	<0.001
	2014	0.40	49.00	17.15	10.29	<0.001
	2015	0.00	76.80	15.65	8.63	<0.001
	2017	6.40	48.60	22.85	6.79	<0.001

Tabelul 1: Indicatori statistici descriptivi pentru parametrii de cantitate şi calitate a laptelui studiaţi şi p-valorile corespunzătoare testului Shapiro-Wilk de verificare a normalităţii

2.4 Modalităţi de analiză a corelaţiei

Pentru analiza asocierii liniare între două variabile cantitative X, Y cu observaţii independente se utilizează în mod uzual coeficientul de corelaţie Pearson, definit prin

$$r_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

unde x_i, y_i reprezintă valorile măsurate pentru X, respectiv Y pentru observaţia i din eşantion, iar n reprezintă volumul eşantionului. Valorile coeficientului de corelaţie Pearson se găsesc în intervalul [-1,1], având următoarea interpretare:

- dacă $r_{XY} < 0$ atunci se poate considera că între X şi Y există corelaţie liniară negativă;

- dacă $r_{XY} \approx 0$ atunci nu există corelaţie liniară între mărimile X şi Y (ceea ce nu înseamnă ca nu există un alt tip de corelaţie);

- dacă $r_{XY} > 0$ atunci se poate considera că între X şi Y există corelaţie liniară pozitivă.

Valorile apropiate de +1, respectiv -1, ale coeficientului indică o corelaţie liniară pozitivă, respectiv negativă, puternică.

Pentru datele din prezentul studiu, dat fiind faptul că acestea reprezintă *măsurători repetate de mai multe ori pe parcursul unui an pentru fiecare animal*, ipoteza de independenţă a observaţiilor nu este satisfăcută. Prin urmare, coeficientul de corelaţie Pearson nu este indicat pentru a investiga gradul de asociere între variabile. În plus, existenţa mai multor seturi de măsurători per animal, deci a mai multor nivele de grupare în date (la nivel de individ vs. populaţie), ridică o problemă cunoscută în statistică drept *paradoxul lui Simpson*, aceea că tendinţele şi asocierile vizibile la un anumit nivel al analizei ar putea fi în contradicţie cu acelea observate la un alt nivel. Figura 7 exemplifică grafic

acest paradox: corelaţia dintre cele două caracteristici figurate X şi Y este pozitivă la nivelul general al populaţiei, dar negativă la nivelul fiecăreia din cele 3 subpopulaţii componente.

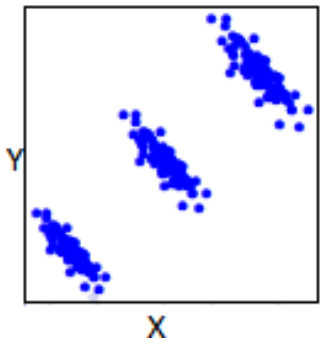


Figura 7: Ilustrare grafică a paradoxului lui Simpson

O alternativă la coeficientul de corelaţie Pearson ce poate fi utilizată în cazul existenţei mai multor măsurători ale aceloraşi variabile pentru fiecare individ este coeficientul de corelaţie pentru măsurători repetate, introdus de Bland şi Altman (Bland & Altman, 1995a), (Bland & Altman, 1995b). Acesta cuantifică asocierea liniară comună la nivel intra-individual a două variabile. Valorile coeficientului de corelaţie pentru măsurători repetate se situează în intervalul [-1,1], interpretarea semnului şi magnitudinii acestora fiind similară cu cea descrisă pentru coeficientul de corelaţie Pearson.

Mai specific, în cazul în care pentru fiecare unitate (subiect) cele două variabile sunt măsurate de mai multe ori, condiţia de independenţă nu mai este satisfăcută. Analiza de asociere adecvată pentru astfel de date depinde de întrebarea la care se doreşte a se răspunde:

- dacă ceea ce se doreşte să se analizeze dacă subiecţii cu valori mari pentru X vor avea de asemenea valori mari pentru Y, se utilizează coeficientul de corelaţie între mediile subiecţilor calculat prin:

$$\frac{n\sum_{i=1}^{k} n_i \bar{x}_i \bar{y}_i - \sum_{i=1}^{k} n_i \bar{x}_i \cdot \sum_{i=1}^{k} n_i \bar{y}_i}{\sqrt{\left(n \cdot \sum_{i=1}^{k} n_i \bar{x}_i{}^2 - (\sum_{i=1}^{k} n_i \bar{x}_i)^2\right) \cdot \left(n \cdot \sum_{i=1}^{k} n_i \bar{y}_i{}^2 - (\sum_{i=1}^{k} n_i \bar{y}_i)^2\right)}},$$

  unde n este numărul total de observaţii, k este numărul de subiecţi, n_i este numărul de observaţii pentru subiectul i, iar $\bar{x}_i$ şi $\bar{y}_i$ reprezintă media lui X, respectiv Y, pentru subiectul i;

- dacă ceea ce se doreşte să se analizeze dacă, pentru un individ, o creştere a valorii lui X este asociată cu creşterea/scăderea valorii lui Y, se va calcula coeficientul de corelaţie pentru măsurători repetate (*rmcorr*) introdus de Bland şi Altman (Bland & Altman, 1995a).

Pentru determinarea lui *rmcorr* se defineşte un model ANOVA care partiţionează variabilitatea lui Y (SS_{tot}) în variabilitate inter-subiecţi (SS_B) şi variabilitate intra-subiect (SS_W). Se consideră apoi un model de regresie liniară cu variabila Y ca răspuns şi variabilele X respectiv subiecţii ca variabile predictive. Astfel, variabilitatea intra-subiect

SS_W determinată de ANOVA se partiţionează din nou, în variabilitate explicată de X (SS_X) şi variabilitate reziduală (SS_{rez}).

In aceste ipoteze se defineşte coeficientul de corelaţie pentru măsurători repetate ca fiind

$$rmcorr = \sqrt{\frac{SS_X}{SS_X + SS_{rez}}},$$

având acelaşi semn cu coeficientul de regresie al lui X din modelul de regresie. Valoarea lui *rmcorr* nu se modifică dacă în analiza de mai sus se interschimbă X cu Y.

La fel ca şi coeficientul de corelaţie Pearson *r*, *rmcorr* ia valori în intervalul $[-1,1]$ şi arată direcţia şi intensitatea *asocierii liniare* dintre cele două variabile (la nivel de individ). În plus *rmcorr* este invariant la transformări liniare asupra lui X şi Y (translaţii sau scalări). Atunci când relaţia dintre X şi Y este diferită la nivel de subiecţi, valoarea lui *rmcorr* va fi aproape de 0. De asemenea o valoare a lui *rmcorr* apropiată de 0 se va obţine atunci când la nivel de subiect, X şi Y nu sunt asociate liniar. Este important de menţionat că *rmcorr* cuantifică asocierea liniară dintre variabile, prin urmare el nu este util în cazul în care la nivel de subiect această asociere este neliniară. Mai multe detalii cu privire la această tehnică de analiză statistică pot fi găsite în (Bakdash&Marusich, 2017).

În Tabelul 2 sunt prezentaţi coeficienţii de corelaţie pentru măsurători repetate dintre *ncs* şi ceilalţi parametri de cantitate şi calitate a laptelui. Analiza acestor corelaţii este relevantă pentru că arată, la nivel de individ, în ce fel apariţia unei infecţii reflectată în creşterea numărului de celule somatice influenţează celelalte caracteristici ale laptelui.

An	*Cantitate*	*Grăsime*	*Lactoză*	*Ph*	*Proteine*	*Raport grăsime / proteine*	*SUN*	*uree*
2012	-0.106 (-0.221, 0.012)	0.108 (-0.009, 0.223)	*-0.379* (-0.476, -0.274)	-0.277 (-0.399 , -0.146)	0.056 (-0.063, 0.172)	0.083 (-0.035, 0.199)	-0.165 (-0.277, -0.048)	0.005 (-0.143, 0.151)
2013	-0.06 (-0.136, 0.010)	0.109 (0.041, 0.178)	*-0.346* (-0.406, -0.284)	-0.073 (-0.153, 0.008)	0.074 (0.004, 0.142)	0.064 (-0.005, 0.133)	-0.045 (-0.114, 0.025)	-0.055 (-0.135, 0.027)
2014	0.05 (-0.033, 0.124)	0.021 (-0.058, 0.100)	-0.144 (-0.221, -0.066)	-0.034 (-0.113, - 0.045)	-0.062 (-0.140, -0.0168)	0.052 (-0.027, 0.131)	-0.141 (-0.217 , -0.062)	-0.052 (-0.130, -0.027)
2015	-0.072 (-0.130, -0.014)	0.138 (0.080, 0.194)	*-0.261* (-0.314, -0.206)	-0.136 (-0.192, -0.078)	0.115 (0.057, 0.172)	0.073 (0.015, 0.130)	0.016 (-0.042, 0.074)	0.007 (-0.051, 0.065)
2017	-0.172 (-0.242, -0.101)	0.188 (0.117, 0.257)	*-0.433* (-0.491, -0.372)	-0.255 (-0.322, -0.186)	0.150 (0.078, 0.220)	0.082 (0.010, 0.154)	0.001 (-0.071, 0.074)	0.014 (-0.059, 0.086)

Tabelul 2: Coeficienţi de corelaţie pentru măsurători repetate dintre *NCS (numărul de celule somatice)* şi caracteristicile producţiei de lapte (*Cantitate, Grăsime, Lactoză, Ph, Proteine,*

Raport Grăsima/ Proteine, SUN, respectiv *Uree)*, pe fiecare an de studiu (valori estimate şi intervale de încredere de 95%)

Se observă corelaţii negative semnificative între *NCS* şi *lactoză* pentru toţi anii investigaţi, precum şi corelaţii foarte slabe sau necorelare cu ceilalţi parametri.

O direcţie viitoare de analiză mai aprofundată a acestor relaţii o reprezintă modelele liniare de regresie multi-nivel (Kreft&Leeuw, 1998).

2.5. Instrumente software utilizate

Prelucrările de statistică descriptivă au fost efectuate în programul R, versiunea 3.4.0. Pentru reprezentările grafice a fost folosit pachetul *ggplot2*, versiunea 3.1.0. În R, coeficientul de corelaţie între mediile subiecţilor se poate calcula folosind funcţia *weightedCorr* din pachetul *wCorr* (Emad&Bailey, 2017), iar coeficientul de corelaţie pentru măsurători repetate se calculează cu funcţia *rmcorr* din pachetul *rmcorr* (Bakdash&Marusich, 2018). Analiza de corelaţie pentru măsurători repetate a fost realizată folosind facilităţile pachetului *rmcorr*, versiunea 0.3.0. Acest pachet conţine suplimentar facilităţi pentru vizualizarea grafică a relaţiei dintre cele două variabile la nivel individual.

3. Analiza datelor genomice

Având în vedere potenţialul pe care îl reprezintă analiza datelor genetice în investigarea rezistenţei antimicrobiene, în prima etapă a proiectului s-a iniţiat documentarea privind studiile de asociere şi familiarizarea cu instrumentele software specifice. Această secţiune a raportului conţine un scurt sumar al informaţiilor extrase din studiul bibliografic respectiv obţinute prin investigarea pachetelor software.

3.1. Specificul studiilor de asociere

Polimorfismul la nivel de nucleotidă (*Single Nucleotide Polimorphism* = SNP = "*snip*") reprezintă cea mai frecventă cauză genetică a variabilităţii dintre indivizi. Cel mai adesea modificarea constă în înlocuirea unei nucleotide (de exemplu citozina cu timina) în secvenţa ADN. Astfel de modificări apar în medie la câte 300 de nucleotide, cel mai adesea în zonele inter-genice, neavând impact asupra sănătăţii sau dezvoltării individului. Snip-urile pot fi interpretate ca fiind markeri biologici iar identificarea lor permit cercetătorilor să localizeze genele asociate cu anumite boli. Când snip-urile apar în cadrul unei gene sau în regiunea din vecinătatea unei gene pot afecta funcţia genei şi influenţa declanşarea unei boli. Cercetătorii au observat că snip-urile pot fi folosite în efectuarea de predicţii privind răspunsul unui individ la anumite medicamente, susceptibilitatea la anumiţi factori de mediu precum şi riscul de a dezvolta anumite boli. Snip-urile pot fi utilizare şi pentru a urmări transmiterea bolilor pe cale genetică.

Scopul studiilor de asociere la nivel de genom (*Genome Wide Association Studies* = GWAS) este de a determina poziţia snip-urilor responsabile de modificări fenotipice şi de a înţelege legătura dintre modificările la nivel de genom şi cele fenotipice.

Etapele principale in procesul de analiză a asocierilor sunt (Wang et al., 2018):

- *Pregătirea datelor.* Datele obţinute prin procesul de genotipare trebuie pregătite şi formatate astfel încât instrumentele software care vor fi utilizate în analiză (de exemplu, pachetul R) să le poată procesa
- *Controlul de calitate.* Este o etapă de filtrare pe baza căreia sunt selectate informaţiile relevante. Criteriile de filtrare utilizate sunt:
  a. *Frecvenţa snip*-ului (call rate) – reprezintă proporţia de snip-uri care au fost genotipate; o valoare egală cu 0.95 semnifică faptul ca 5% din valori sunt absente; în general snip-urile cu un call rate mai mic decât 1 sunt ignorate
  b. *Frecvenţa alelei minore* (Minor –allele frequency = MAF) – reprezintă frecvenţa celei mai puţin comune alele pentru fiecare snip; snip-urile cu MAF mai mic decât un prag (de exemplu 0.05) sunt ignorate
  c. *Analiza exploratorie datelor.* In această etapă se aplică tehnici din data mining cu scopul de a investiga relaţiile dintre subiect. Una dintre cele mai populare tehnici este cea de analiză a componentelor principale ("Principal Component Analysis"=PCA).
- *Aplicarea modelelor statistice.* In contextul GWAS cele mai frecvent utilizate modele sunt cele mixte care ţin cont de stratificarea populaţiei şi de relaţiile existente între teste de de asociere. Pentru a testa dacă un anumit snip are influenţa asupra unui fenotip specific se poate folosi un model de forma $y = X\beta + Zg + S\tau + \varepsilon$ unde β este vectorul efectelor fixate (modelează variabilele fenotipice), g modelează caracteristicile genetice (ca efect aleator, cu o anumită variabilitate $Var[g] = K\sigma^2$) iar τ modelează efectul snip-ului. Eroarea modelului are varianţa $Var[\varepsilon] = I\sigma_e^2$.

3.2. Instrumente software specifice

Unul dintre instrumentele software cele mai folosite pentru studii de asociere este platforma pentru prelucrări statistice R. Pe lângă funcţiile de bază care pot fi utilizate pentru pre-procesarea datelor există pachete specifice analizei snip-urilor. Etapa corespunzătoare controlului de calitatea poate fi realizată folosind pachetul *HapEstXXR (Multilocus Stepwise Regression)* care combină regresia de tip stepwise cu analiza bazată pe haplotipuri. De exemplu pentru a realiza filtrarea datelor pe baza frecvenţei alelei minore se poate folosi funcţia *maf*. Pentru a utiliza această funcţie datele trebuie pregătite sub formă tabelară: liniile conţin informaţii despre subiecţii incluşi în analiză, iar coloanele corespund snip-urilor (cu excepţia primei coloane care conţine identificatorii subiecţilor). Datele obţinute în urma genotipării sunt de regulă specificate sub forma (exemple preluate din (Mota et al, 2018)):

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
[1,]	B-1	..	AB	AB	BB	AA	AB	AA	AB	BB
[2,]	B-2	AB	AA	AA	BB	AA	BB	AA	BB	BB
[3,]	B-3	AA	AA	..	BB	..	AB	AA	AA	BB
[4,]	B-4	AB	AA	AB	AB	AB	AA	BB	BB	AB
[5,]	B-5	..	AB	AB	AA	..	..	..	AA	AA
[6,]	B-6	BB	AA	AB	AB	AA	BB	BB	AB	AB
[7,]	B-7	AA	BB	..	AA	AB	..	AA	AA	AA
[8,]	B-8	AB	AB	BB	BB	AA	BB	BB	AB	AB
[9,]	B-9	AB	AA	AB	AA	..	..	AA	AB	BB
[10,]	B-10	BB	BB	AA	BB	BB	BB	BB	AB	AB

Intrucât funcţia maf prelucrează date numerice, tabelele de tipul de mai sus trebuie transformate folosind regulile următoare:

- Valorile absente se specifică prin 0
- „AA” se înlocuieşte cu 1
- „BB” se înlocuieşte cu 2
- „AB” se înlocuieşte cu 3.

Prin aplicarea acestei transformări se obţine:

	id	rs11102647	rs6695241	rs12567796	rs2810583	rs4654986	rs1567710	rs1320964	rs2205847	rs10923099
[1,]	B-1	0	3	3	2	1	3	1	3	2
[2,]	B-2	3	1	1	2	1	2	1	2	2
[3,]	B-3	1	1	0	2	0	3	1	1	2
[4,]	B-4	3	1	3	3	3	1	2	2	3
[5,]	B-5	0	3	3	1	0	0	0	1	1
[6,]	B-6	2	1	3	3	1	2	2	3	3
[7,]	B-7	1	2	0	1	3	0	1	1	1
[8,]	B-8	3	3	2	2	1	2	2	3	3
[9,]	B-9	3	1	3	1	0	0	1	3	2
[10,]	B-10	2	2	1	2	2	2	2	3	3

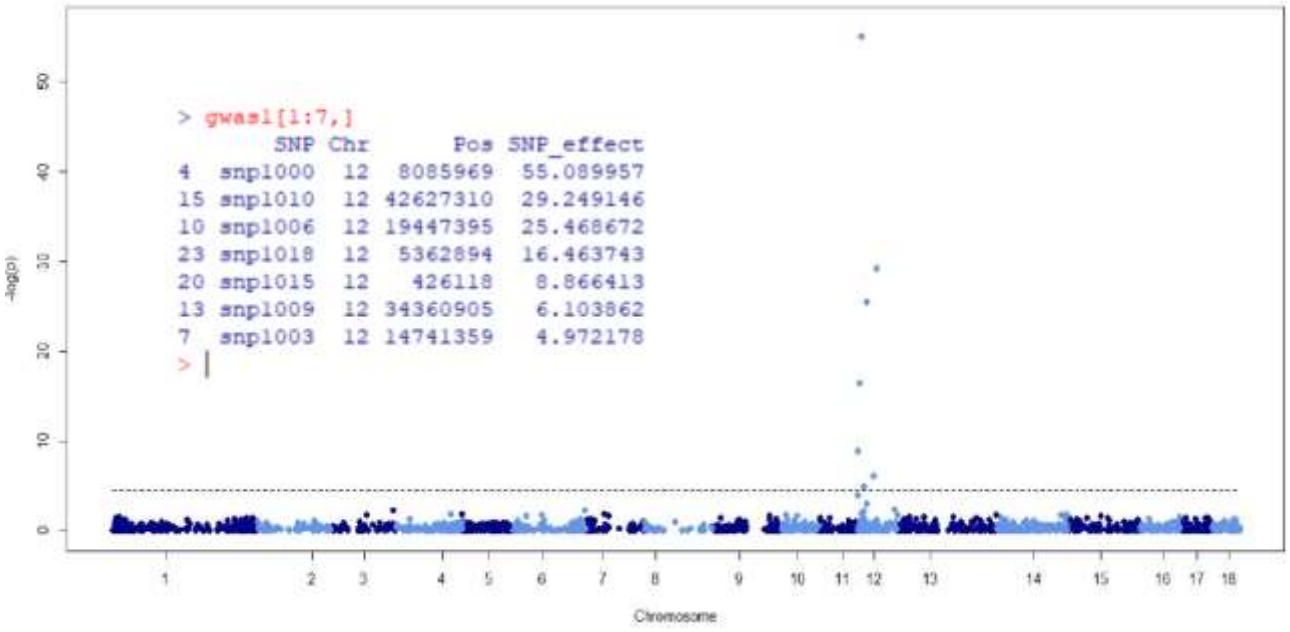
Funcţia *maf* aplicată unui set de date având structura de mai sus (folosind sintaxa de apel *maf(dataset, marker.label = snp_names)* unde *dataset* reprezintă tabelul de date, iar snp_names este o listă cu numele snip-urilor*)* returnează un set de rezultate cu structura următoare:

	X1.1	X1.2	X2.2	Total	NMISS	call.rate	minor.allele	maf	hwe.chisq.p.value
snp1	53	196	188	437	63	0.874	1	0.3455378	0.861660336
snp2	70	177	196	443	57	0.886	1	0.3577878	0.005993345
snp3	68	197	173	438	62	0.876	1	0.3801370	0.339816094
snp4	63	189	187	439	61	0.878	1	0.3587699	0.177915467
snp5	70	179	202	451	49	0.902	1	0.3536585	0.005112940
snp6	66	169	202	437	63	0.874	1	0.3443936	0.002683199

In tabelul de mai sus coloanele *X1.1*, *X1.2* şi *X2.2* conţin numărul de snip-uri de fiecare tip, *NMISS* conţine numărul de valori absente, *call.rate* conţine raportul *Total/(Total+NMISS)* (proporţia de snip-uri prezente) iar maf reprezintă frecvenţa alelei minore (alela notată cu 1) corespunzătoare fiecărui snip, adică raportul *(X1.1+X1.2)/Total*. Valorile astfel calculată pot fi utilizate în etapa de filtrare (de exemplu, se reţin doar snip-urile pentru care *call.rate>0.95* şi *maf>0.05).* După selecţia snip-urilor se procedează la analiza statistică folosind pachetul rrBLUP. De exemplu prin apelul

gwas=GWAS(pheno, geno, fixed=NULL, n.PC=0, plot=TRUE)

unde *pheno* reprezintă setul de date ce conţine caracteristicile fenotipice iar geno conţine caracteristicile genotipice se obţine:



Bibliografie:

(Bakdash&Marusich, 2017) J. Z. Bakdash, L. R. Marusich – Repeated Measures Correlation, Frontiers in Psychology 8 (2017), Article 456, doi: 10.3389/fpsyg.2017.00456

(Bakdash&Marusich, 2018) Bakdash, J. Z. and Marusich, L. R., rmcorr: Repeated Measures Correlation. R package version 0.3.0 (2018). https://CRAN.R-project.org/package=rmcorr

(Bland&Altman, 1995a) J. M. Bland, D. G. Altman – Calculating correlation coefficients with repeated observations: part 1 Correlation within subjects, BMJ 310 (1995), 446. doi:10.1136/bmj.310.6977.446

(Bland&Altman, 1995b) J. M. Bland, D. G. Altman – Calculating correlation coefficients with repeated observations: part 2 Correlation between subjects, BMJ 310 (1995), 633. doi: 10.1136/bmj.310.6980.633

(Emad&Bailey, 2017) Emad, A. and Bailey, P., wCorr: Weighted Correlations. R package version 1.9.1 (2017). https://CRAN.R-project.org/package=wCorr

(Mota et al, 2018) R.R. Mota, S. Naderi, H. Hammami, N. Gengler, Modeling and GWAS, Genotype plus Environment Workshop, Bucharest, 2018

(Kreft&Leeuw, 1998) I. Kreft, J. de Leeuw – Introducing Multilevel Modeling. Thousand Oaks, CA, SAGE Publications, 1998.

(Wang et al., 2018) M.H. Wang, H.J. Cordell, K. Van Steen, Statistical methods for genome-wide association studies, Seminars in Cancer Biology, 2018.

Anexa 1. Extras din codul R utilizat în preprocesarea datelor – pregătirea pentru analiza exploratorie; codul a fost adaptat pentru specificul datelor din fiecare an

```
######## Prelucrare fisierelor referitoare la cantitatea de lapte
cant=read.table(file.choose(), sep='\t')
cant1= cant[,c(1,2,6,8,9,10,11,12,14,15,16,18,20,21)] # extragerea coloanelor relevante
colnames(cant1)=c("ID","c1","c2","c3","c4","c5","c6","c7","c8","c9","c10","c11","c12","c13")
c3_num=rep(3, length(cant1[,3]))
cant2= cbind(cant1, c3_num)
cant1$c1_num=1
cant1$c2_num=2
cant1$c10_num=10
c1_num=rep(1,length(cant1[,1]))
c2_num=rep(2,length(cant1[,2]))
c3_num=rep(3,length(cant1[,3]))
c4_num=rep(4,length(cant1[,4]))
c5_num=rep(5,length(cant1[,5]))
c6_num=rep(6,length(cant1[,6]))
c7_num=rep(7,length(cant1[,7]))
c8_num=rep(8,length(cant1[,8]))
c9_num=rep(9,length(cant1[,9]))
c10_num=rep(10,length(cant1[,10]))
c11_num=rep(11,length(cant1[,11]))
c12_num=rep(12,length(cant1[,12]))
c13_num=rep(13,length(cant1[,13]))
cant2=cbind(cant1,c1_num,c2_num,c3_num,c4_num,c5_num,c6_num,c7_num,c8_num,c9_nu
m,c10_num,c11_num,c12_num,c13_num)

data1=cant2[,c("ID","c1_num","c1")]
colnames(data1)=c("ID","measure","cant")
data2=cant2[,c("ID","c2_num","c2")]
colnames(data2)=c("ID","measure","cant")
data3=cant2[,c("ID","c3_num","c3")]
colnames(data3)=c("ID","measure","cant")
data4=cant2[,c("ID","c4_num","c4")]
colnames(data4)=c("ID","measure","cant")
data5=cant2[,c("ID","c5_num","c5")]
colnames(data5)=c("ID","measure","cant")
data6=cant2[,c("ID","c6_num","c6")]
colnames(data6)=c("ID","measure","cant")
data7=cant2[,c("ID","c7_num","c7")]
colnames(data7)=c("ID","measure","cant")
data8=cant2[,c("ID","c8_num","c8")]
colnames(data8)=c("ID","measure","cant")
data9=cant2[,c("ID","c9_num","c9")]
colnames(data9)=c("ID","measure","cant")
data10=cant2[,c("ID","c10_num","c10")]
colnames(data10)=c("ID","measure","cant")
```

```
data11=cant2[,c("ID","c11_num","c11")]
colnames(data11)=c("ID","measure","cant")

data12=cant2[,c("ID","c12_num","c12")]
colnames(data12)=c("ID","measure","cant")
data13=cant2[,c("ID","c13_num","c13")]
colnames(data13)=c("ID","measure","cant")

# concatenarea liniilor cu valorile corespunzatoare tuturor masuratorilor
cant_total= rbind
(data1,data2,data3,data4,data5,data6,data7,data8,data9,data10,data11,data12,data13)

######## Cazeina
caze=read.table(file.choose(), sep='\t')
caze1= caze[,c(1,3,7,9,11,14,17,20,22,25,27,29,31,33)] # extragerea coloanelor relevante
colnames(caze1)=c("ID","c1","c2","c3","c4","c5","c6","c7","c8","c9","c10","c11","c12","c13")
c1_num=rep(1,length(caze1[,1]))
c2_num=rep(2,length(caze1[,2]))
c3_num=rep(3,length(caze1[,3]))
c4_num=rep(4,length(caze1[,4]))
c5_num=rep(5,length(caze1[,5]))
c6_num=rep(6,length(caze1[,6]))
c7_num=rep(7,length(caze1[,7]))
c8_num=rep(8,length(caze1[,8]))
c9_num=rep(9,length(caze1[,9]))
c10_num=rep(10,length(caze1[,10]))
c11_num=rep(11,length(caze1[,11]))
c12_num=rep(12,length(caze1[,12]))
c13_num=rep(13,length(caze1[,13]))

caze2=cbind(caze1,c1_num,c2_num,c3_num,c4_num,c5_num,c6_num,c7_num,c8_num,c9_nu
m,c10_num,c11_num,c12_num,c13_num)
data1=caze2[,c("ID","c1_num","c1")]
colnames(data1)=c("ID","measure","caze")
data2=caze2[,c("ID","c2_num","c2")]
colnames(data2)=c("ID","measure","caze")
data3=caze2[,c("ID","c3_num","c3")]
colnames(data3)=c("ID","measure","caze")
data4=caze2[,c("ID","c4_num","c4")]
colnames(data4)=c("ID","measure","caze")
data5=caze2[,c("ID","c5_num","c5")]
colnames(data5)=c("ID","measure","caze")
data6=caze2[,c("ID","c6_num","c6")]
colnames(data6)=c("ID","measure","caze")
data7=caze2[,c("ID","c7_num","c7")]
colnames(data7)=c("ID","measure","caze")
data8=caze2[,c("ID","c8_num","c8")]
colnames(data8)=c("ID","measure","caze")
```

```
data9=caze2[,c("ID","c9_num","c9")]
colnames(data9)=c("ID","measure","caze")

data10=caze2[,c("ID","c10_num","c10")]
colnames(data10)=c("ID","measure","caze")
data11=caze2[,c("ID","c11_num","c11")]
colnames(data11)=c("ID","measure","caze")
data12=caze2[,c("ID","c12_num","c12")]
colnames(data12)=c("ID","measure","caze")
data13=caze2[,c("ID","c13_num","c13")]
colnames(data13)=c("ID","measure","caze")
caze_total=
rbind(data1,data2,data3,data4,data5,data6,data7,data8,data9,data10,data11,data12,data13
)

######## Prelucrari similare pentru celelalte marimi de interes: grasime, lactoza, ncs, Ph,
proteine, raport grasime/proteine, SUN, uree.

[……]


######## Construirea fisierului global corespunzator unui an

require(plyr) # pachet pentru agregarea datelor
library(plyr)

total <- join_all(list(cant_total, caze_total, gras_total, lact_total, ncs_total, ph_total,
prot_total,gras_prot_total,sun_total,uree_total), by = c("ID","measure"), type = 'full')

write.table(total, "../BioAMR/Data/2017/total_2017.txt", sep="\t")
```