Razvoj i validacija metoda vibracijske spektroskopije za identifikaciju pročišćenih meningokoknih polisaharida serogrupa A i C

Mandac Zubak, Ana

Sveučilište u Zagrebu

Prehrambeno-biotehnološki fakultet

Ana Mandac Zubak

RAZVOJ I VALIDACIJA METODA VIBRACIJSKE SPEKTROSKOPIJE ZA IDENTIFIKACIJU PROČIŠĆENIH MENINGOKOKNIH POLISAHARIDA SEROGRUPA A I C

DOKTORSKI RAD

Mentor:

Prof. dr. sc. Anita Slavica

Zagreb, 2021.

University of Zagreb

Faculty of Food Technology and Biotechnology

Ana Mandac Zubak

DEVELOPMENT AND VALIDATION OF VIBRATIONAL SPECTROSCOPY METHODS FOR IDENTIFICATION OF PURIFIED MENINGOCOCCAL POLYSACCHARIDES SEROGROUPS A AND C

DOCTORAL DISERTATION

Supervisor:

Ph.D. Anita Slavica, Full Professor

Zagreb, 2021.

Tema doktorskog rada pod naslovom „Razvoj i validacija metoda vibracijske spektroskopije za identifikaciju pročišćenih meningokoknih polisaharida serogrupa A i C“ prihvaćena je na sjednici Fakultetskog Vijeća Prehrambeno-biotehnološkog fakulteta Sveučilišta u Zagrebu održanoj 28. listopada 2020. godine, a Senat Sveučilišta u Zagrebu donio je odluku o pokretanju postupka stjecanja doktorata znanosti na sjednici održanoj 15. prosinca 2020.

Informacije o mentoru

Prof. dr. sc. Anita Slavica, Prehrambeno-biotehnološki fakultet Sveučilišta u Zagrebu

Prof. dr. sc. Anita Slavica rođena je u Šibeniku 08. svibnja 1970. godine. Po završetku srednje škole upisuje Prehrambeno-biotehnološki fakultet Sveučilišta u Zagrebu (PBF SuZ), na kojem je diplomirala Biokemijsko inženjerstvo. Od 1996. radi u Zavodu za biokemijsko inženjerstvo (Zavod za BI) PBF SuZ, gdje sudjeluje u nastavi, znanstveno-istraživačkom radu i pohađa poslijediplomski znanstveni magistarski studij Biotehnologija. Magistrirala je 2002. godine iz područja Biotehničkih znanosti, znanstveno polje Biotehnologija. U ak. god. 2002./2003. upisuje poslijediplomski znanstveni doktorski studij *Technical Chemistry* na *Graz University of Technology*. Paralelno na *Faculty for Chemistry, Chemical- and Process Engineering, and Biotechnology* i *Research Centre for Applied Biocatalysis* sudjeluje u istraživanjima na više međunarodnih znanstveno-istraživačkih projekata, a primarno na projektu *Determination and improvement of operational stability of enzymes*, kojeg podupire tvrtka Sandoz. Ovdje usvaja cijeli niz novih znanja i vještina iz područja molekularne biotehnologije i bioprocesnog inženjerstva, kao i pravila znanstveno-istraživačkog rada i upravljanja i zaštite intelektualnog vlasništva. Sve predmete propisane doktorskim studijem *Technical Chemistry* je položila s prosječnom ocjenom 1,2 (4,8) i svoj doktorski rad obranila 2006. godine mit *Auszeichnung bestandene* na *Graz University of Technology* (https://online.tugraz.at). Time je stekla akademski stupanj *Doktorin der technichen Wissenschaften* (Doctor technicae, Dr. techn.). Rješenjem Agencije za znanost i visoko obrazovanje Republike Hrvatske je stekla akademski stupanj doktor tehničkih znanosti (dr. sc.). Nakon toga, nastavlja svoj znanstveno-istraživački i nastavni rad u Zavodu za BI, gdje je 2017. izabrana za redovitu profesoricu. Nositelj je i suradnik na nekoliko predmeta na preddiplomskim i diplomskim studijskim programima; poslijediplomskom sveučilišnom doktorskom studiju Biotehnologija i bioprocesno inženjerstvo, prehrambena tehnologija i nutricionizam; i sveučilišnom interdisciplinarnom poslijediplomskom specijalističkom studiju Intelektualno vlasništvo (http://intvla.unizg.hr). Uvela je najnovije sadržaje i priredila nastavne materijale za nekoliko predmeta, koji se izvode na SuZ. Odlukom Matičnog odbora za biotehničke znanosti od 27. veljače 2019. izabrana je u znanstveno zvanje znanstveni savjetnik u trajnom zvanju.

Znanstvenim istraživanjima, inovativnim tehnologijama i transferom tehnologija bavila se i bavi se u okviru nacionalnih i međunarodnih znanstveno-istraživačkih projekta, jednim *Proof*

of Concept i jednim nacionalnim stručnim projektom. Od 2018. do danas nositeljica je potpore istraživanjima primjene robusnih biokatalizatora u održivoj biotehnološkoj proizvodnji biokemikalija i drugih proizvoda visoke dodane vrijednosti na PBF SuZ. Objavila je 20 znanstvenih radova, autorica je jednog patenta (HR P20100074 A2) i suradnica na tekstu iz područja biotehnologije u Hrvatskoj tehničkoj enciklopediji. Uređuje međunarodne znanstvene časopise i recenzirala je više od 50 znanstvenih radova, organizirala više od 20 međunarodnih znanstvenih skupova, sudjelovala na više od 30 međunarodnih znanstvenih skupova i predsjedala sesijama na šest međunarodnih znanstvenih kongresa. Održala je niz plenarnih i pozvanih predavanja na međunarodnim i nacionalnim znanstvenim skupovima kao i devet predavanja na *Graz University of Technology*. Mentor je osam završnih radova, 25 diplomskih radova i tri doktorska rada. Dobitnica je nekolika nagrada u zemlji i inozemstvu.

(Su)Osnivačica, predsjednica, dopredsjednica i članica je nekoliko znanstvenih međunarodnih i nacionalnih organizacija i stručnih društava kao i različitih Odbora na PBF SuZ. Kroz različita predstavnička tijela zastupa interese Republike Hrvatske pri tijelima Europske komisije, prenosi znanstveno utemeljene činjenice iz područja biotehnologije u široj društvenoj zajednici i zastupa kreiranje boljeg, pravednijeg i prosperitetnijeg društva.

Najljepše zahvaljujem mentorici prof. dr. sc. Aniti Slavici na znanstvenom dijalogu, razumijevanju, uloženom trudu i vremenu te pomoći tijekom izrade ovog doktorskog rada.

Hvala kolegama u Laboratoriju za kemijsku kontrolu kvalitete Imunološkog zavoda u Zagrebu na poticajnoj znanstveno istraživačkoj i radnoj atmosferi.

Hvala prof. dr. sc. Jadranki Frece na iznimnoj podršci i potpori pri realizaciji ovog doktorskog rada.

Veliko hvala prof. dr. sc. Jagodi Šušković i prof. dr. sc. Blaženki Kos na neiscrpnoj podršci i konstruktivnim i korisnim savjetima tijekom doktorskog studija.

Zahvaljujem se kolegama iz Pliva Hrvatska, kao i kolegama sa Medicinskog fakulteta Sveučilišta u Zagrebu na pomoći pruženoj prilikom eksperimentalne faze rada.

Hvala Ines Fabijanić na moralnoj podršci, potpori i motivaciji svih ovih godina na putu prema doktoratu znanosti.

Hvala svima koji su vjerovali u mene.

Za Maru i Unu!

TEMELJNA DOKUMENTACIJSKA KARTICA

Sveučilište u Zagrebu **Doktorski rad**
Prehrambeno-biotehnološki fakultet
Sveučilišni poslijediplomski (doktorski) studij Biotehnologija i bioprocesno inženjerstvo
UDK: 543.424.2: 543.635.25: 66.085.1(043.3)
Znanstveno područje: Biotehničke znanosti
Znanstveno polje: Biotehnologija

RAZVOJ I VALIDACIJA METODA VIBRACIJSKE SPEKTROSKOPIJE ZA IDENTIFIKACIJU PROČIŠĆENIH MENINGOKOKNIH POLISAHARIDA SEROGRUPA A I C

Ana Mandac Zubak, dipl. ing. biotehnologije

Rad je izrađen u Imunološkom zavodu; Prehrambeno-biotehnološkom fakultetu Sveučilišta u Zagrebu; Pliva Hrvatska, Zagreb; te na Medicinskom fakultetu Sveučilišta u Zagrebu.

Mentor: Prof. dr. sc. Anita Slavica

Kratki sažetak

Cilj ovog rada je bio istražiti mogućnost primjene novih metoda vibracijske spektroskopije - spektroskopije bliskoga infracrvenoga zračenja (NIR) i Ramanove spektroskopije, za identifikaciju pročišćenih meningokoknih polisaharida serogrupa (PMPS) A i C. Primjenom kemometrije formirani su NIR i Raman SIMCA i PLS-DA modeli. Optimizirani i validirani modeli imaju 100 %-tnu učinkovitost klasificiranja nepoznatih uzoraka PMPS A i C. Oba NIR modela, koji su formirani na temelju podataka o proizvodnim serijama PMPS A i C, su učinkovito potvrdili identitet standarda polisaharida A i C. Pristupom NIR i Raman SIMCA jednoklasnog modeliranja utvrđena je 100 %-tna učinkovitost autentifikacijskih modela za, zasebno, PMPS A i PMPS C. Primjena NIR i Raman modela za identifikaciju PMPS A i C je izvrsna alternativa referentnoj dvostrukoj imunodifuziji - Ouchterlony metodi u uporabi. Kvalitativni istraživački pristup i rezultati ovog rada mogu se primijeniti u razvoju novih NIR i Raman modela za identifikaciju drugih meningokoknih polisaharida kao i za analizu složenih matriksa i klasifikacijsku primjenu, kako u farmaceutskoj tako i u biotehnološkoj industriji.

Broj stranica: 232
Broj slika: 238
Broj tablica: 15
Broj literaturnih navoda: 109
Jezik izvornika: hrvatski

Ključne riječi: meningokokni polisaharidi serogrupa A i C, NIR spektroskopija, PCA, PLS-DA, Raman spektroskopija, SIMCA

Datum obrane: 09. srpnja 2021.

Stručno povjerenstvo za obranu:

1. Prof. dr. sc. Jagoda Šušković
2. Prof. dr. sc. Višnja Gaurina Srček
3. Dr. sc. Marin Roje, viši znanstveni suradnik
4. Prof. dr. sc. Blaženka Kos (zamjena)

Rad je pohranjen u knjižnici Prehrambeno-biotehnološkog fakulteta u Zagrebu, Kačićeva 23 i u Nacionalnoj i sveučilišnoj knjižnici u Zagrebu, Hrvatske bratske zajednice 4, te na Sveučilištu u Zagrebu, Trg Republike Hrvatske 14, Zagreb.

BASIC DOCUMENTATION CARD

University of Zagreb **Ph. D. Thesis**
Faculty of Food Technology and Biotechnology
Postgraduate university (doctoral) study Biotechnology and Bioprocess Engineering
UDK: 543.424.2: 543.635.25: 66.085.1(043.3)
Scientific Area: Biotechnical sciences
Scientific Field: Biotechnology

DEVELOPMENT AND VALIDATION OF VIBRATIONAL SPECTROSCOPY METHODS FOR IDENTIFICATION OF PURIFIED MENINGOCOCCAL POLYSACCHARIDES SEROGROUPS A AND C

Ana Mandac Zubak, dipl. ing. biotechnology

Thesis was performed at Institute of Immunology, Zagreb; Faculty of Food Technology and Biotechnology University of Zagreb; Pliva Croatia, Zagreb; and at School of Medicine, University of Zagreb.
Supervisor: Ph.D. Anita Slavica, Full Professor

Short abstract

The aim of this study was to investigate the possibility of applying new methods of vibrational spectroscopy - near infrared spectroscopy (NIR) and Raman spectroscopy, to identify purified meningococcal polysaccharides serogroups (PMPS) A and C. By the use of chemometrics NIR and Raman SIMCA and PLS-DA models were formed. Optimized and validated models have 100 % efficiency in classifying unknown samples of PMPS A and C. Both NIR models, which were formed based on PMPS A and C production series spectral data, effectively confirmed the identity of polysaccharide standards A and C. 100 % efficient authentication models for PMPS A and PMPS C, separately, have been obtained by using NIR and Raman with the SIMCA single-class modeling approach. Application of NIR and Raman models for identification of PMPS A and C is an excellent alternative to the reference double immunodiffusion - Ouchterlony method in use. Qualitative research approach and the results of this work can be implemented in the development of new NIR and Raman models for the identification of other meningococcal polysaccharides as well as for the analysis of complex matrices and classification application, both in the pharmaceutical and biotechnology industries.

Number of pages: 232
Number of figures: 238
Number of tables: 15
Number of references: 109
Original in: Croatian

Keywords: meningococcal polysaccharides serogroups A and C, NIR spectroscopy, PCA, PLS-DA, Raman spectroscopy, SIMCA

Date of the thesis defense: July 09th 2021

Reviewers:
1. Ph. D. Jagoda Šušković, Full professor
2. Ph. D. Višnja Gaurina Srček, Full professor
3. Ph. D. Marin Roje, Senior research associate
4. Ph. D. Blaženka Kos, Full professor (substitute)

Thesis is deposited in: Library of Faculty of Food Technology and Biotechnology University of Zagreb, Kačićeva 23; National and University Library, Hrvatske bratske zajednice 4; and also at University of Zagreb, Trg Republike Hrvatske 14, Zagreb.

SAŽETAK

Pročišćeni meningokokni polisaharidi (PMPS) A i C su djelatne tvari cjepiva protiv meningokoka. Potvrda identiteta svakog PMPS prvi je korak u kontroli kvalitete proizvodnje ovog cjepiva. Europska farmakopeja propisuje identifikaciju meningokoknih polisaharida refrentnom dvostukom imunodifuzijom - Ouchterlony metodom. Iako je ova referentna metoda vrlo specifična, njezina uporaba povlači nekoliko vrlo značajnih nedostataka, kako slijedi. Ouchterlony metoda podrazumijeva složene i dugotrajne postupke, ekonomski je zahtjevna i, glede uzorka, destruktivna. U izvedbi ove referentne metode koristi se antiserum životinjskog podrijetla, a nije zanemariva uporaba drugih kemikalija i potrošnog materijala, tako da je i ekološka strana ove metode zaista nepovoljna. Naime, Direktiva 2010/63/EU Europskog parlamenta i Vijeća od 22. rujna 2010. upućuje na zamjenu, redukciju i usavršavanje uporabe kemikalija i drugog materijala životinjskog podrijetla kao i razvoj i primjenu novih, naprednih analitičkih metoda, u kojima se ne koriste materijali životinjskog podrijetla.

Brojna istraživanja metoda za identifikaciju različitih polisaharida upućuju da se metode vibracijske spektroskopije - spektroskopija bliskoga infracrvenoga zračenja (NIR) i Ramanova spektroskopija, u kombinaciji s kemometrijom, mogu koristiti za razvoj novih, brzih i robusnih modela za identifikaciju i klasifikaciju polisaharida. U usporedbi s konvencionalnim analitičkim tehnikama, metode vibracijske spektroskopije imaju brojne prednosti, kao što su: nije potrebna zasebna procedura za pripravu uzoraka, metode nisu destruktivne, podaci se brzo prikupljaju, svaki uzorak se može nanovo analizirati, a prikupljeni i spremljeni spektri se brzo analiziraju i uspoređuju. Ove su napredne metode sasvim u skladu s Direktivom 2010/63/EU.

Na temelju dostupnih znanstvenih podataka može se sa sigurnošću ustvrditi da identifikacija meningokoknih polisaharida A i C pomoću NIR i Ramanove spektroskopije nije do sada nikada provedena i ovaj je doktorski rad prvi takav napor u primjeni novih metoda vibracijske spektroskopije u identifikaciji PMPS A i C.

Jedan od ciljeva ovog rada je bio snimiti NIR i Raman spektre proizvodnih serija PMPS A i C kao i spektre eksperimentalnih serija PMPS W135 i Y (koji su bili negativne probe, a čija je kemijska struktura slična kemijskoj strukturi PMPS C), a koje su proizvedene u Imunološkom zavodu u Zagrebu. Kao prvi korak u multivarijatnoj analizi eksperimentalnih podataka primijenjene su različite matematičke metode predobrade snimljenih spektara uz korištenje programa Unscrambler v 10.4 (Camo Analytics AS, Oslo, Norveška). Načinjena je eksploracijska analiza glavnih komponenti (PCA) spektara obrađenih različitim matematičkim obradama gdje su grupirani slični i jasno razdvojeni različiti spektralni podaci PMPS A i C.

Profilom opterećenja identificirana su najvažnija spektralna područja, koja definiraju glavne komponente (PC) i, shodno tome, diferencijaciju ovih meningokoknih polisaharida.

Provedeno je meko neovisno modeliranje analogne klase (SIMCA) i određen je broj glavnih komponenti. Optimizacija formiranog NIR SIMCA modela provela se test setom 1, dok se optimizacija formiranog Raman SIMCA modela provela postupkom unakrsne validacije. U uspješnoj optimizaciji ovih SIMCA modela korišteni su uzorci PMPS A, C, W135 i Y u cilju provjere specifičnosti obaju modela. NIR SIMCA model je uspješno klasificirao i standarde PMPS A i PMPS C. Klasifikacijska sposobnost NIR i Raman SIMCA modela utvrđena je na temelju dobivenih validacijskih parametara (osjetljivosti, specifičnosti i učinkovitosti), kako za pojedinu klasu tako i za cjelokupne modele. Validacija NIR i Raman SIMCA modela provedena je vanjskim setom uzoraka PMPS A, C, W135 i Y, koji nije korišten u kalibraciji niti u optimizaciji dvaju SIMCA modela. Utvrđena je valjanost formiranih i validiranih NIR i Raman SIMCA modela, koji uspješno klasificiraju 100 % nepoznatih uzoraka PMPS A i C i razlikuju ne ciljne skupine PMPS - W135 i Y. Učinkovit je bio i pristup jednoklasne klasifikacije NIR i Raman SIMCA modela, gdje su ciljni uzorci PMPS A ili C, zasebno, uspješno identificirani, dok su preostali ne ciljni uzorci (PMPS W135 i Y) ostali neklasificirani.

Diskriminantnom analizom parcijalnih najmanjih kvadrata (PLS-DA) formirani su pripadajući NIR i Raman modeli. Učinkovita optimizacija NIR PLS-DA modela provedena je test setom 1 načinjenim od PMPS A, C, W135 i Y te standarda PMPS A i C. Očekivano (zbog sličnosti njihove kemijske strukture) NIR PLS-DA model je identificirao PMPS W135 i Y kao PMPS C. Obzirom da uzorci PMPS W135 i Y nisu u redovnoj proizvodnji, nema mogućnosti pogrešnih identifikacijskih slučajeva u rutinskoj primjeni PLS-DA modela za identifikaciju PMPS A i C. Također, NIR PLS-DA model uspješno je klasificirao standarde PMPS A i C. Kao i kod SIMCA modela, optimizacija Raman PLS-DA modela provedena je postupkom unakrsne validacije. Validacija NIR i Raman PLS-DA modela načinjena je kao i kod SIMCA modela.

Uspješna primjena NIR i Raman modela za identifikaciju PMPS A i C izvrsna je i ekonomski učinkovita alternativa Ouchterlony metodi u uporabi. Kvalitativni istraživački pristup i rezultati opisani u ovom doktorskom radu mogu se primijeniti u razvoju novih NIR i Raman modela za identifikaciju drugih polisaharida. Ovdje opisani rezultati upućuju na potencijal metoda vibracijske spektroskopije u kombinaciji sa multivarijantnim tehnikama za analizu složenih bioloških matriksa u farmaceutskoj i biotehnološkoj industriji općenito, osobito u autentifikacijske svrhe.

Ključne riječi: meningokokni polisaharidi serogrupa A i C, NIR spektroskopija, PCA, PLS-DA, Raman spektroskopija, SIMCA

ABSTRACT

Purified meningococcal polysaccharides serogroups (PMPS) A and C are the active pharmaceutical substances of the meningococcal vaccine. Confirmation of the identity of each PMPS is the first step in quality control of production of this vaccine. The European Pharmacopoeia prescribes the identification of meningococcal polysaccharides by the reference double immunodiffusion - Ouchterlony method. Although this reference method is very specific, its use entails several very significant drawbacks, as follows. Ouchterlony method involves complex and time-consuming procedures, it is economically challenging, and sample destructive. The antiserum of animal origin is used in the performance of this reference method, and the use of other chemicals and consumables is not negligible, so the ecological side of this method is really unfavorable. Namely, according to the Directive 2010/63/EU of the European Parliament and of the Council on 22 September 2010 the replacement, reduction and refinement of the use of chemicals and other animal materials as well as the development and application of new, advanced analytical methods, that do not use animal materials origin, should be implemented.

Numerous studies of methods for the identification of different polysaccharides suggest that vibrational spectroscopy methods - near-infrared spectroscopy (NIR) and Raman spectroscopy, combined with chemometry, can be used to develop new, fast and robust models for the identification and classification of polysaccharides. Compared to conventional analytical techniques, vibration spectroscopy methods have a number of advantages, such as: no separate procedure for sample preparation is required, methods are not destructive, data are acqired quickly, each sample can be re-analyzed, and collected and stored spectra can be quickly analyzed and compared. These advanced methods are fully in line with the Directive 2010/63/EU.

Based on available scientific data, it can be stated that the identification of meningococcal polysaccharides A and C by NIR and Raman spectroscopy has never been performed and this doctoral thesis is the first such effort in applying new vibrational spectroscopy methods in the identification of PMPS A and C.

One of the aims of this work was to acquire the NIR and Raman spectra of the production series of PMPS A and C as well as the spectra of the experimental series of PMPS W135 and Y (which were negative samples and whose chemical structure is similar to the chemical structure of PMPS C), all produced at the Institute of Immunology in Zagreb. Principal component analysis (PCA) of mathematicaly processed spectra (the program Unscrambler v 10.4; Camo Analytics AS, Oslo, Norway) were applied, where similar and clearly separated different spectral data of

PMPS A and C were grouped. The most important spectral regions were identified by the loading profile, which define the main components (PC) and, consequently, differentiating meningococcal polysaccharides.

Soft independent modeling of class analogy (SIMCA) was performed and a number of major components were determined. The optimization of the formed NIR SIMCA model was performed by test set 1, while the optimization of the formed Raman SIMCA model was performed by the cross-validation procedure. In the successful optimization of these SIMCA models, samples PMPS A, C, W135 and Y were used in order to check the specificity of both models. The NIR SIMCA model has successfully classified both the PMPS A and PMPS C standards. The classification capability of the NIR and Raman SIMCA models was determined based on the obtained validation parameters (sensitivity, specificity and efficiency), both for each class and for the entire models. Validation of the NIR and Raman SIMCA models was performed with an external set of PMPS samples A, C, W135 and Y, which was not used in the calibration or optimization of the two SIMCA models. The efficacy of formed and validated NIR and Raman SIMCA models, which successfully classify 100 % of unknown PMPS samples A and C and distinguish non-target groups PMPS - W135 and Y, was determined. The approach of one-class classification of NIR and Raman SIMCA models was effective, where the target samples are PMPS A or C, separately, were successfully identified, while the remaining non-target samples (PMPS W135 and Y) remained unclassified.

The corresponding NIR and Raman models were formed by partial least squares discriminant analysis (PLS-DA). Effective optimization of the NIR PLS-DA model was performed by test set 1 made of PMPS A, C, W135 and Y and PMPS A and C standards. As expected (due to the similarity of their chemical structure) the NIR PLS-DA model identified PMPS W135 and Y as PMPS C. Since the PMPS W135 and Y samples are not in regular production, there is no possibility of incorrect identification in the routine application of the PLS-DA model to identify PMPS A and C. Also, the NIR PLS-DA model has successfully classified the PMPS A and C standards. Again, as in the SIMCA model, the optimization of the Raman PLS-DA model was performed by the cross-validation procedure. Validation of NIR and Raman PLS-DA models was done in the same way as with SIMCA models.

Successful application of the NIR and Raman models to identify PMPS A and C is an excellent and cost-effective alternative to the Ouchterlony method in use. The qualitative research approach and results described in this doctoral thesis can be applied in the development of new NIR and Raman models for the identification of other polysaccharides. The results described here point out the potential of vibrational spectroscopy methods in combination with

multivariate techniques for the analysis of complex biological matrices in the pharmaceutical and biotechnology industries in general, in particular for authentication purposes.

Keywords: meningococcal polysaccharides serogroups A and C, NIR spectroscopy, PCA, PLS-DA, Raman spectroscopy, SIMCA

SADRŽAJ

1. UVOD

Invazivne meningokokne bolesti predstavljaju globalni javnozdravstveni problem i cijepljenje se smatra najučinkovitijim načinom prevencije ovih bolesti, osobito kod male djece. Kao i kod drugih farmaceutskih proizvoda, kompleksna, zahtjevna i dugotrajna biotehnološka proizvodnja kao i osiguranje kvalitete i sigurnosti cjepiva podliježe strogim regulatornim zahtjevima i podrazumijeva opsežnu analitičku karakterizaciju (među)proizvoda kroz sve faze proizvodnje do stavljanja krajnjeg proizvoda na tržište (Ph. Eur.2019; USP 40-NF 35, 2017; WHO TRS 594, 1975). Procjenjuje se da oko 70 % trajanja ukupnog procesa proizvodnje cjepiva pripada kontroli kvalitete ovih izuzetno važnih biotehnoloških proizvoda. Složenost i dugotrajnost regulatorno uvjetovanih analitičkih metoda odgađaju puštanje cjepiva na tržište i to uz negativne zdravstvene ali i ekonomske implikacije te, posljedično, gubitak konkurentnosti tvrtke koja proizvodi cjepiva. Iz svega navedenog slijedi da je jedan od glavnih imperativa za prevladavanje ovih zdravstvenih i ekonomskih poteškoća razvoj bržih, jeftinijih i ekološki prihvatljivih analitičkih metoda kontrole kvalitete (među)proizvoda farmaceutske ali i općenito biotehnološke industrije. Nadalje, zakonski propisi za zaštitu životinja koje se koriste u znanstvene svrhe (Direktiva 2010/63/EU Europskog parlamenta i Vijeća od 22. rujna 2010.) pripadaju skupini najstrožih etičkih standarda koji se primjenjuju diljem svijeta, a koji su na snazi u zemljama Europske unije od 01. siječnja 2013. Ova direktiva upućuje na potrebu primjene tzv. principa "tri r" (od engl. replacement, reduction and refinement) ili principa zamjene, redukcije i usavršavanja uporabe kemikalija i drugog materijala životinjskog podrijetla kao i razvoj i primjenu alternativnih materijala i novih, naprednih analitičkih metoda, u kojima se ne koriste materijali životinjskog podrijetla.

Znanstveno-istraživačka zajednica prepoznaje i primjenjuje metode vibracijske spektroskopije - spektroskopiju bliskog infracrvenog zračenja (NIR) i Raman spektroskopiju, kao napredne, ne destruktivne, brze i jednostavne metode, kod kojih je potreban relativno mali uzorak za ove preferirane jeftine analize (Willett i Rodriguez, 2018; Rodionova i sur., 2019; Rodionova i Pomerantsev, 2020). Kod razvoja novih analitičkih metoda i njihove primjene u kontroli kvalitete cjepiva, naglasak se stavlja i na validaciju ovih metoda te primjenu novih statističkih i matematičkih alata (Lopez i sur., 2015; Riedl i sur., 2015; Brereton i sur, 2017; Ermer, 2015). Cjepivo protiv meningokoka, koje od 1978. godine proizvodi Imunološki zavod, sastoji se od pročišćenih meningokoknih kapsularnih polisaharida serogrupa (PMPS) A i C. Identifikacija ovih djelatnih tvari (PMPS A i C) pripada među ključne postupke u proizvodnji cjepiva protiv meningokoka. Zahtjev 21 CFR 610.14 Agencije za hranu i lijekove, SAD (CFR, 2019) propisuje identifikaciju i potvrdu identiteta sadržaja spremnika iz svake serije proizvedenoga cjepiva. Identifikacija meningokoknih polisaharida refrentnom dvostukom imunodifuzijom -

Ouchterlony metodom (Ouchterlony, 1962) propisana je Europskom farmakopejom (Ph. Eur. 2019). Međutim, ova refrentna imunokemijska metoda visoke specifičnosti je destruktivna, dugotrajna, ekonomski zahtjevna i uključuje uporabu antiseruma životinjskog podrijetla.

Stoga je cilj ovoga doktorskog rada bio istražiti mogućnost primjene NIR i Raman spektroskopije uz primjenu različitih kemometrijskih alata kao alternativnih postupaka za identifikaciju PMPS A i C. Razvojem i validacijom novih valjanih NIR i Raman modela nedvojbeno bi se identificirali nepoznati uzorci PMPS A i C. Tako bi se referentna Ouchterlony metoda mogla uspješno zamijeniti s novim metodama vibracijske spektroskopije s izuzetnim potencijalom u rutinskoj primjeni kontrole kvalitete djelatnih tvari i cjepiva.

2. OPĆI DIO

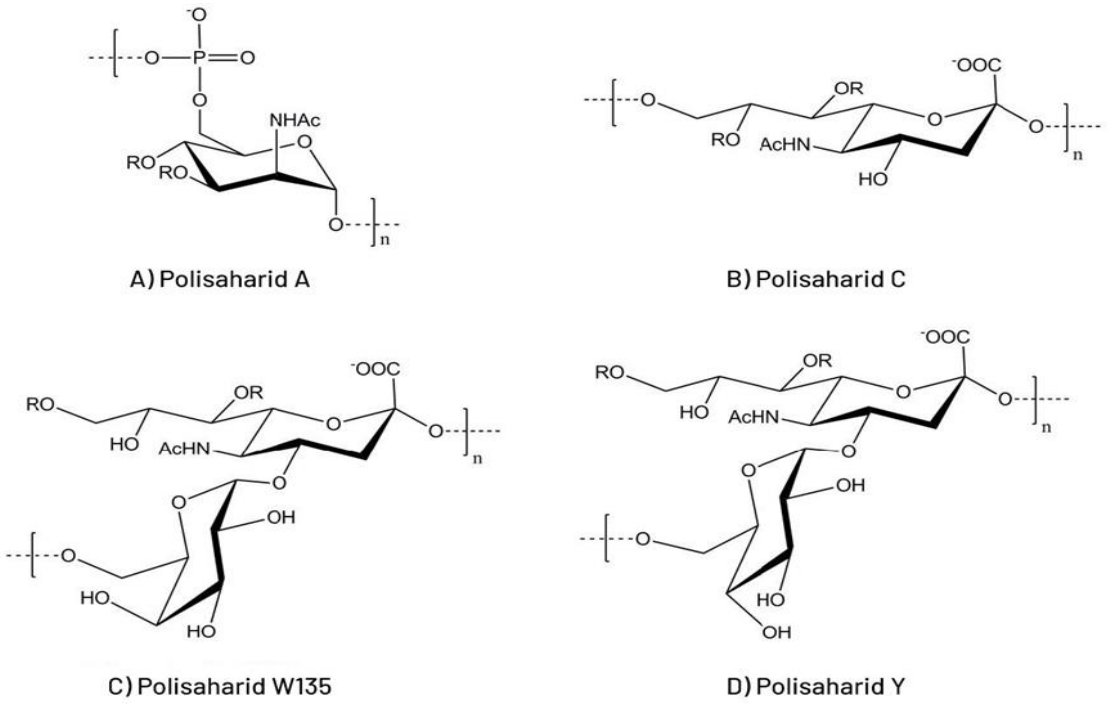
2.1. Meningokokno cjepivo

Meningokokna bolest još uvijek predstavlja globalni zdravstveni problem (WHO, 1999). Etiološki uzročnik meningokoknog meningitisa je *Neisseria meningitidis* (također se naziva *Meningococcus*), Gram-negativna bakterija, po prvi put izolirana davne 1887. godine (Weichselbaum, 1887) iz likvora bolesnika oboljelog od meningitisa. Meningokoki ne predstavljaju antigenski homogenu skupinu. Poznato je najmanje 13 serogrupa meningokoka s različito eksprimiranim kapsularnim polisaharidima, koji su dalje klasificirani u serotipove i pod-serotipove u skladu sa strukturom kapsularnih polisaharida. Epidemiološki najvažnije serogrupe koje uzrokuju većinu meningokoknih bolesti su serogrupe: A, B, C, W135 i Y, a nedavno je i okarakterizirana i serogrupa X (Gabutti i sur., 2015).

Meningokoki izazivaju bolest u ljudi, a prirodno im je boravište sluznica gornjih dišnih putova. Prenose se s osobe na osobu kapljičnim putem te direktnim kontaktom s respiratornim sekretom.

Cijepljenje se smatra najboljim načinom prevencije meningokokne bolesti, osobito kod dojenčadi i male djece (Panatto i sur., 2013). Razvoj cjepiva mijenjao se tijekom vremena, kako je opisano ovdje dalje. Prva cjepiva protiv meningokoka, koja su razvijena od 1900-te do 1940-te, a koja su sadržavala inaktiviranu cjelostaničnu suspenziju *N. meningitidis*, nisu bila uspješna i to zbog visoke reaktogenosti i nemogućnosti da se jasno procijeni njihova zaštitna uloga, a zbog upotrebe neadekvatne eksperimentalne kontrole cjepiva. Daljnji razvoj ovog cjepiva usmjerio se na specifične serogrupe kapsularnih polisaharida *N. meningitidis*. Scherp and Rake su 1945. godine kod miševa dokazali zaštitnu ulogu seruma imuniziranih konja (Scherp i Rake, 1945). Nažalost, ti pročišćeni kapsularni polisaharidi nisu inducirali imunološki odgovor kod ljudi, što je bilo povezano s neodgovarajućim postupkom izolacije, koji je rezultirao uglavnom polisaharidima male molekulske mase. Kasnije je otkriveno da su za indukciju imunološkog odgovora potrebni polisaharidi velike molekulske mase (Kabat i sur., 1944; Kabat i Bezer, 1958). Stoga su, 1960-ih Gotschlich i sur. (1969a, 1969b) razvili alternativni pristup za pročišćavanje polisaharida velike molekulske mase, što je bila osnova za proizvodnju novog cjepiva protiv meningokoka serogrupa A i C. Prvo cjepivo protiv meningokoka serogrupe C registrirano je u SAD-u 1974. godine, dok je kombinirano cjepivo (koje se sastoji od serogrupa A, C, W135 i Y; Slika 1.) registrirano 1978 godine. U Hrvatskoj je razvoj meningokoknog polisaharidnog cjepiva započeo u Imunološkom zavodu u Zagrebu ranih 1970-ih. Početni rad na proizvodnji cjepiva temeljio se na radovima Gotschlich i sur. (1969a, 1969b) te na vlastitim istraživanjima u svezi fizikalno-kemijskih i imunogenih karakteristika meningokoknih

polisaharida sergrupa A i C. Prvo su proizvedena monovalentna cjepiva tj. cjepivo protiv meningokoka serogrupe A i cjepivo protiv meningokoka serogrupe C, a zatim kombinirano cjepivo protiv meningokoka serogrupa A i C, koje je proizvedeno tijekom 1977. godine i prvi puta registrirano 1978. godine.



Slika 1. Struktura meningokoknih polisaharida serogrupa A, C, W135 i Y u obliku homopolimera (serogrupe A i C) i heteropolimera (serogrupe W135 i Y). NHAc - N-acetilne grupe; R-(O)Ac ili (O)H; OAc - O-acetilne grupe.

Meningokokni polisaharid sergrupe A je homopolimer, koji se sastoji od djelomično *O*-acetiliranih N-acetilmanozamin monomerskih jedinica povezanih u položaju α(1→6) fosfodiesterskim vezama. To je kapsularni polisaharid koji zajedno s endotoksinom stanične stjenke *N. meningitidis* predstavlja najvažniji virulentni faktor ove bakterije. Meningokokni polisaharid grupe C je homopolimer, koji se sastoji od monomernih jedinica koje čine djelomično *O*-acetilirana *N*-acetilneuraminska kiselina (sijalinska kiselina). Monomerne jedinice povezane su glikozidnim vezama u položaju α(2→9). Nativni polisaharid osnovni je sastojak kapsule *N. meningitidis* serogrupe C i zajedno s lipopolisahardiom (endotoksinom stanične stijenke ove bakterije) smatra se glavnim virulentnim faktorom. Meningokokni polisaharid Y je heteropolimer građen od djelomično *O*-acetilirane *N*-acetilneuraminske kiseline (sijalinska kiselina) i D-glukoze, vezanih α(2→6) i α(1→4) glikozidnim vezama.

Polisaharid W135 je heteropolimer građen naizmjenično od djelomično *O*-acetilirane *N*-acetilneuraminske kiseline (sijalinska kiselina) i D-galaktoze, veznih α(2→6) i α(1→4) glikozidnim vezama (Slika 1).

2.2. Identifikacija materijala u postupku kontrole kvalitete biotehnološke proizvodnje meningokoknog cjepiva

U farmaceutskoj se industriji moraju identificirati i odobriti svi materijali (ekscipijenti, pomoćne tvari, djelatne tvari te gotovi proizvodi) u upotrebi, a koji odgovaraju određenoj namjeni. Cilj identifikacije ovih materijala je potvrda identiteta ciljne supstancije. Identifikacija je stoga prvi korak u složenom postupku kontrole kvalitete biotehnološke proizvodnje cjepiva. Ponekad je teško povući granicu između identifikacije i karakterizacije kemijske strukture ili sastava (među)proizvoda, na što Europska farmakopeja daje odgovor: „Ispitivanja navedena u odjeljku Identifikacija nisu dizajnirana da pruže potpunu potvrdu kemijske strukture ili sastava proizvoda; ona su namijenjena potvrdi, s prihvatljivim stupnjem sigurnosti da je proizvod u skladu s opisom na naljepnici.“ (Ph.Eur., 2019; Görög, 2015). Monografije u farmakopejama obično počinju s testovima za identifikaciju, jer je jednostavna ali pouzdana identifikacija od velike važnosti u izbjegavanju opasnosti i pogrešaka. Testiranje identiteta provodi se u svrhu razlikovanja ispitivanog uzorka od ostalih materijala koji se proizvode na istom mjestu. Potvrda identiteta zahtjeva selektivne tehnike sa visokom specifičnosti. Specifičnost identifikacije bi trebala moći razlikovati aktivne i pomoćne tvari sličnih kemijskih struktura. Također, metode ne smiju biti preosjetljive kako bi se izbjegle lažne reakcije uzrokovane toleriranim nečistoćama, te ne trebaju biti eksperimentalno zahtjevnije od potrebnog razlikovanja predmetne tvari od ostalih dostupnih farmaceutskih tvari. Zahtjevnije instrumentalne metode koje se koriste za potvrdu identiteta su: spektrofotometrijske analize, poput infracrvene spektroskopije (IR) i nuklearne magnetske rezonancije (NMR); te kromatografske metode poput plinske kromatografije (GC) ili tekućinske kromatografije (LC). Ostale metode uključuju: (a) određivanje fizikalnih konstanti, poput: tališta, ledišta, vrelišta, specifične optičke rotacije, kuta rotacije, ultraljubičastog (UV) spektra, specifične apsorbancije, relativne gustoće, refraktivnog indeksa i viskoznosti; zatim (b) kemijskih reakcija, kao što su reakcije obojenja ili taložne reakcije, također određivanje kemijskih vrijednosti (vrijednosti saponifikacije, estera, hidroksila i joda), te (c) tankoslojna tekućinska kromatografija (TLC) (Ermer, 2015).

Sukladno zahtjevu 21 CFR 610.14 Agencije za hranu i lijekove, SAD (CFR, 2019), sadržaj završnog spremnika svakog punjenja i svake serije cjepiva nakon završetka označavanja, testira se na identitet. Prema zahtjevu WHO TRS (WHO, 2014), potrebno je provesti test identiteta pročišćenih polisaharida kako bi se potvrdio njihov identitet. U slučaju kada se na istom proizvodnom mjestu proizvode i drugi polisaharidi, metodu je potrebno validirati kako bi se sa sigurnošću utvrdilo da razlikuje željeni polisaharid od svih ostalih polisaharida proizvedenih na toj proizvodnoj lokaciji. Prema preporuci Europske faramakopeje (Ph.Eur., 2019) identifikacija meningokoknih polisaharida i meningokoknog cjepiva se provodi metodom po Ouchterlony-u.

2.3. Metoda dvostruke imunodifuzije (Ouchterlony metoda)

Dvostruka imunodifuzija (Ouchterlony metoda) koristi se za kvalitativnu procjenu specifičnog vezanja antigen-antitijelo. Metoda se zasniva na reakciji antigena i specifičnog antitijela u sloju agara, pri čemu nastaje precipitacijska vrpca. Reakcija se provodi u sloju agar i agaroza gela, koji se koriste kao polučvrsti mediji, čineći proces difuzije postojanim te omogućavaju uočavanje precipitacijskih linija.

U dvostrukoj imunodifuziji (Ouchterlony), koncetracijski se gradijent uspostavlja za oba reaktanta, i za antigen i za antitijelo. Antigen i antitijelo difundiraju jedan prema drugome i na mjestu gdje koncentracija difundiranih antitijela i antigena postigne točku ekvivalencije, nastaje ravna linija precipitata. Budući da molekule veće molekulske mase difundiraju znatno sporije od molekula manje molekulske mase, položaj precipitacijske linije je djelomično funkcija molekulske mase obaju reaktanata - antigena i antitijela.

Početne koncentracije antigena i antitijela obično su podešene tako da se linija precipitacije formira na pola puta između dvaju zdenaca. Međutim, ukoliko su antitijela u suvišku, linija će nastati bliže zdencu antigena, a ako je u suvišku antigen, linija će biti bliža zdencu antitijela.

Prije započinjanja izvođenja dvostruke difuzijske tehnike u kvalitativne svrhe, nužno je optimirati uvjete za ovu precipitaciju. Ukoliko se radi s tzv. neuravnoteženim sustavom, npr. ukoliko postoji suvišak antigena ili antitijela, može doći do putovanja precipitata ili pojave višestrukog precipitata i to zbog naizmjeničnog otapanja i ponovne precipitacije kompleksa antigen - antitijelo. Kako bi se odredio optimalni odnos koncentracije antigena i antitijela, koji je nužan za precipitaciju, antiserum se nanese u centralni zdenac, a serija razrijeđenja antigena nanosi se u zdence koji okružuju centralni zdenac (Slike 2.).

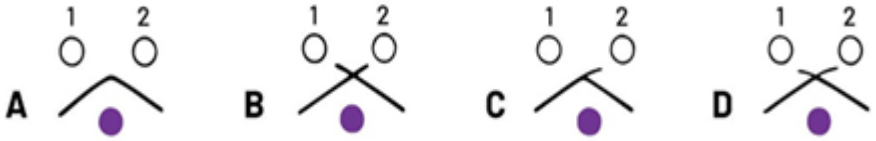
2.3.1. Oblici precipitacije

U sloju agaroza gela izbuše se tri zdenca; u oba gornja zdenca nanesu se uzorci antigena, a u donji zdenac nanese se antiserum, koji sadrži antitijela za sve determinante antigena. Tri su osnovna oblika precipitacije, koji se mogu postići pomoću ove tehnike (Slika 2.):

A.) Uzorak identiteta: kontinuirani precipitat se formira ako su dva antigena potpuno identična ili imaju zajedničku determinantu. Antitijela u antiserumu reagiraju sa oba antigena. Antitijela ne mogu razlikovati antitigene tj. antigeni su imunološki identični. Ova se reakcija naziva reakcija potpune identičnosti.

B.) Uzorak neidentiteta: dva odvojena precipitata se formiraju ako dva antigena nisu identična ili nemaju niti jedne zajedničke determinante. Niti jedno antitijelo u antiserumu ne reagira sa antigenim determinantama koje mogu biti prisutne u oba antigena tj. antigeni nisu imunološki povezani s antiserumom. Ova se reakcija naziva reakcija neidentičnosti.

C.) i D.) Uzorak djelomičnog identiteta: kontinuirani precipitat s lukom se formira ako jedan antigen ima jednu determinantu suviše i ako ona nije zajednička za oba antigena. Antitijela u antiserumu s jednim od antigena reagiraju više nego s drugim. Ova se reakcija naziva reakcija djelomične identičnosti.

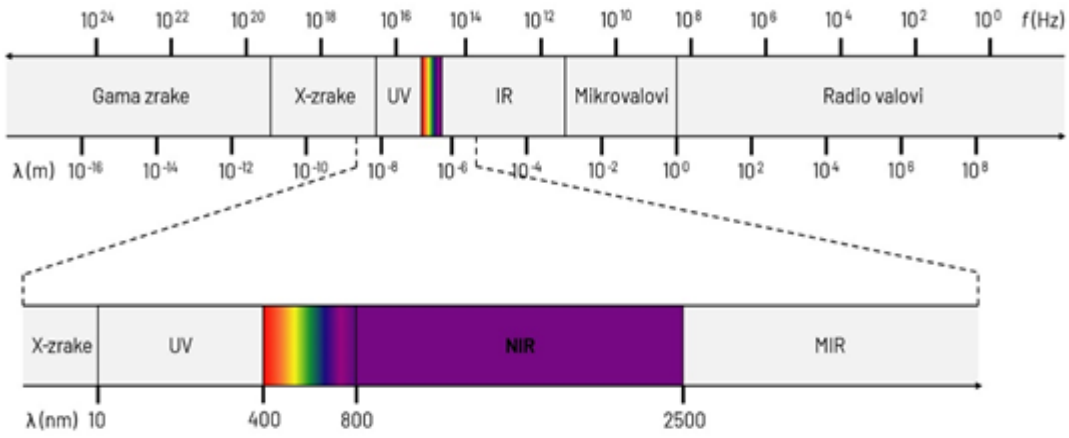


Slika 2. Različiti oblici precipitacije kod Ouchterloney dvostruke imunodifuzije. Gornji zdenci 1 i 2 (prazni kružić) su uzorci antigena, donji zdenac (ljubičasti kružić) predstavlja antiserum, koji sadrži antitijela za sve determinante antigena.

2.4. Vibracijska spektroskopija

Interakcija svjetlosti s materijom obuhvaća čovjekov interes tijekom posljednja dva tisućljeća. Sva svjetlost klasificirana je kao elektromagnetsko (EM) zračenje i sastoji se od izmjeničnih električnih i magnetskih polja, koja su klasično opisana kontinuiranim sinusoidnim valovitim kretanjem električnog i magnetskog polja (Larkin, 2018). Spektroskopija je znanost koja

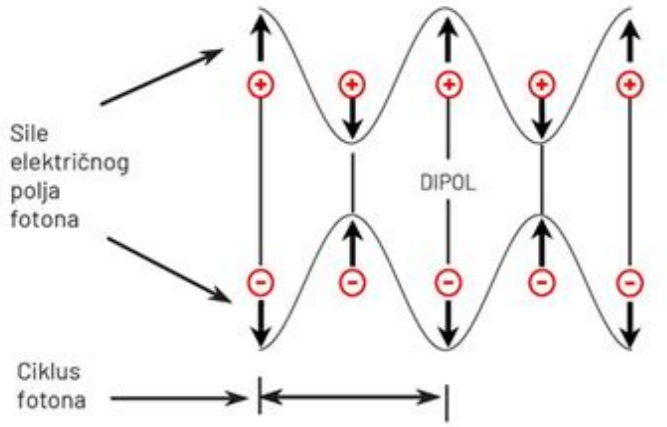
proučava interakciju elektromagnetskog zračenja (Slika 3.) i materije, odnosno bavi se istraživanjima energetskih nivoa u atomima, molekulama, kristalima i sl. (Cid i Bravo, 2015). Za razliku od atoma, molekule uz elektronske energetske nivoe posjeduju dodatne stupnjeve slobode - molekulske vibracije i rotacije. Vibacijska spektroskopija ima vrlo širok spektar primjene i nudi rješenja za mnoštvo važnih i izazovnih analitičkih problema (Larkin, 2018). I IR i Raman spektroskopija su molekularne vibracijske spektroskopske tehnike pomoću kojih se mogu proučavati vibracijski prijelazi u molekulama. Kod IR i Raman spektroskopije uzima se u obzir samo električno polje, a zanemaruje komponenta magnetskog polja. Pomoću IR i Raman spektroskopije mogu se proučavati interakcije EM zračenja s molekularnim vibracijama, ali se međusobno razlikuju u načinu na koji se energija fotona prenosi na molekulu mijenjajući njezino vibracijsko stanje (Larkin, 2018). Dok je Ramanova spektroskopija tehnika raspršivanja EM zračenja , IR spektroskopija se temelji na apsorpciji EM zračenja (Burns, 2008). Budući da su vibracijske energetske razine jedinstvene za svaku molekulu, IR i Raman spektar daju „otisak prsta“ (engl. *fingerprint*) određene molekule (Larkin, 2018). IR i Raman vibracijske tehnike su komplementarne tehnike, odnosno mogu se nadopunjavati. Molekule koje proizvode dobar signal u IR spektru, mogu proizvesti slabi signal u Raman spektru i obrnuto. Raman i IR spektri sadrže kvalitativne i kavantitativne informacije o kemijskom sastavu i fizikalnim svojstvima uzorka (De Beer i sur., 2011).



Slika 3 Spektar bliskog infracrvenog zracenja (NIR) u kontekstu spektra elektromagnetskog zračenja.

2.4.1. Infracrvena spektroskopija

Infracrvena spektroskopija (engl. *infrared spectroscopy*, IR) proučava interakcije EM zračenja i tvari u rasponu bliskog (valni brojevi, $\tilde{\nu}$ = 13333 cm^{-1} - 4000 cm^{-1}), srednjeg ($\tilde{\nu}$ = 4000 cm^{-1} - 400 cm^{-1}) i dalekog ($\tilde{\nu}$ = 400 cm^{-1} - 10 cm^{-1}) područja IR spektra (Larkin, 2018; Pasquini, 2003). IR spektroskopija detektira prijelaze između molekularnih vibracijskih energetskih nivoa kao rezultat apsorpcijskog IR zračenja. Da bi se energija iz IR fotona prenijela na molekulu apsorpcijom, molekularne vibracije moraju uzrokovati promjenu dipolnog momenta molekule. Ova interakcija između svjetlosti i materije je pravilo odabira za IR spektroskopiju koje zahtjeva promjenu dipolnog momenta tijekom vibracije da bi bilo IR aktivno (Slika 4).
IR spektar se dobiva odnosom intenziteta (apsorbancije ili transmisije) prema valnom broju ($\tilde{\nu}$) i ovaj je odnos proporcionalan energetskoj razlici između osnovnog i pobuđenog vibracijskog stanja (Larkin, 2018). Prijelaz molekule iz osnovnog u pobuđeno vibracijsko stanje u spektru se detektira kao vrpca.



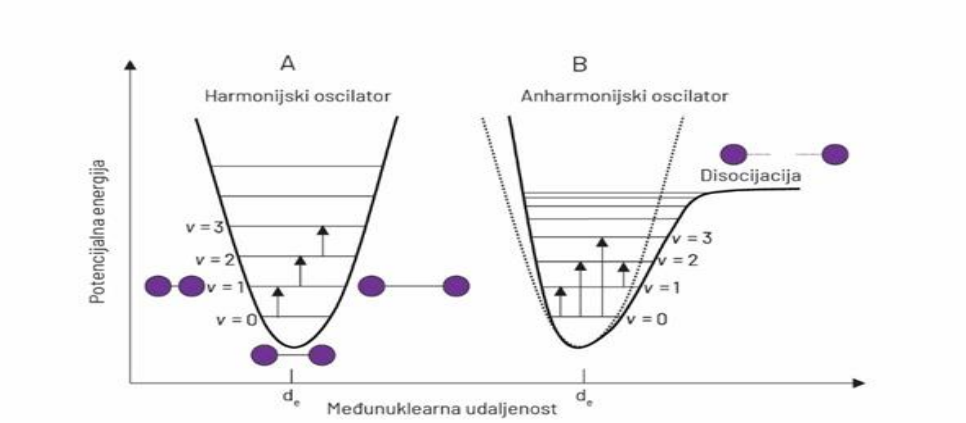
Slika 4. Prikaz oscilirajućeg električnog polja fotona, koje formira oscilirajuće suprotno usmjerene sile na pozitivne i negativne naboje molekulskog dipola (Larkin, 2018).

2.4.1.1.Spektroskopija bliskog infracrvenog zračenja (NIR)

1800. godine engleski astronom Wiliam Harschel (Harschel, 1800) je pokušavajući odgovoriti na pitanje koja boja u vidljivom spektru donosi toplinu od sunca, otkrio NIR zračenje. Staklenom prizmom je razdvojio sunčevu svjetlost a toplomjerom je odredio promjenu temperature svake boje. Zabilježio je lagano povećanje temperature kroz spektar, a učinak

zagrijavanja prema crvenom kraju spektra je postao očit. Slučajno, nakon što je ostavio termometar izvan spektra pored crvene boje, temperatura se znatno povećala. Herschel je ovaj novootkriveni fenomen nazvao "zračenjem topline" i "termometrijskim spektrom". Pogrešno je smatrao da se ovaj oblik energije razlikuje od svjetlosti. Danas se ovi zaključci mogu činiti iznenađujućima, ali tada još nije postojao koncept elektromagnetskog spektra, niti da je vidljiva svjetlost samo njegov mali dio. Harschel je pretpostavio nevidljivu vrpcu izvan crvene ili na latinskom *infra* red (Burn, 2001; Burns, 2008). Vrlo složeni NIR spektri i naizgled beznadni rezultati doveli su do zanemarivanja NIR spektroskopije sve do početka 1970-ih, kada je NIR spektrometar spojen s računalom. Kompleksni NIR spektar sadrži veliki broj informacija koje se mogu iskoristiti za kvalitativne i kvantitativne analitičke svrhe, te je posljednjih godina NIR spektroskopija postala nezamjenjiv alat u znanosti ali i u industrijskoj primjeni (Blanco i Villaroya 2002; Pasquini 2003).

Blisko infracrveno područje (engl. *near infrared*, NIR) obuhvaća elektromagnetsko zračenje u rasponu valnih brojeva $\tilde{\nu}$ = 13333 cm^{-1} - 4000 cm^{-1} (λ = 750 - 2500 nm; Slika 3.) (Pasquini, 2003). Dva su osnovna načina vibriranja molekula - vibracije istezanja (simetrično i asimetrično) i vibracije savijanja (u ravnini i izvan ravnine). Vibracija molekula može se opisati modelom harmonijskog oscilatora (Slika 5.A), pomoću kojeg se može izračunati energija različitih, jednako raspoređenih razina (energije vibracije prikazuju se kao ekvidistantni pravci unutar površine parabole). Mogući su samo prijelazi između uzastopnih energetskih razina koji uzrokuju promjenu dipolnog momenta (Blanco i Villarroya, 2002). Međutim, u praksi takozvani idealni harmonijski oscilator ima ograničenja tj. ne može objasniti ponašanje molekula, jer ne uzima u obzir Coulomb-ovu odbojnost među atomima ili disocijaciju veza u molekulama. Tako se ponašanje molekula može opisati modelom anharmonijskog oscilatora (Slika 5.B), kod kojeg energetske razine nisu jednoliko raspoređene (razlike između susjednih energetskih razina nisu konstantne, već se smanjuju s porastom vibracijskog kvantnog broja).



Slika 5. Shematski prikaz harmonijskog (A) i anharmonijskog (B) oscilatora za potencijalnu energiju dvoatomske molekule. de - ravnotežna udaljenost u molekuli, *v* – energetske razine molekule (Pasquini, 2003).

Intenzitet NIR vrpci ovisi o promjeni dipolnog momenta molekule i anharmoničnosti veze u molekuli. Budući da je atom vodika najlakši i stoga pokazuje najveće vibracije i najveća odstupanja od harmoničnog ponašanja, glavne vrpce tipično uočene u NIR regiji odgovaraju vezama koje sadrže ovaj i druge lake atome, kao što su: C - H, N - H, O - H i S - H. Suprotno tome, vrpce veza poput C = O, C - C i C - Cl mnogo su slabije ili čak odsutne. Također, NIR spektar sadrži ne samo kemijske, već i fizikalne podatke, koje se mogu koristiti za određivanje fizikalnih svojstava uzoraka (Blanco and Villarroya, 2002).
Anharmoničnost može rezultirati prijelazima između energetskih razina $\Delta v = \pm 2, \pm 3, ...$, i ovi prijelazi poznati su kao viši tonovi (engl. *overtons*). Ove vrpce se javljaju između 780 nm i 2000 nm, ovisno o redoslijedu viših tonova i prirodi i snazi veze (Blanco i Villarroya, 2002).Druga važna značajka NIR spektra je veliki broj kombinacijskih vrpci (Burns, 2001). U poliatomskim molekulama dva ili više vibracijskih načina mogu međudjelovati na takav način da uzrokuju istodobne energetske promjene i stvaraju apsorpcijske vrpce, koje se nazivaju kombinacijske vrpce. NIR kombinacijske vrpce javljaju se između 1900 nm i 2500 nm. NIR spektar sadrži apsorpcijske vrpce koje odgovaraju višim tonovima i kombinacijama osnovnih vibracija (Blanco i Villarroya, 2002).

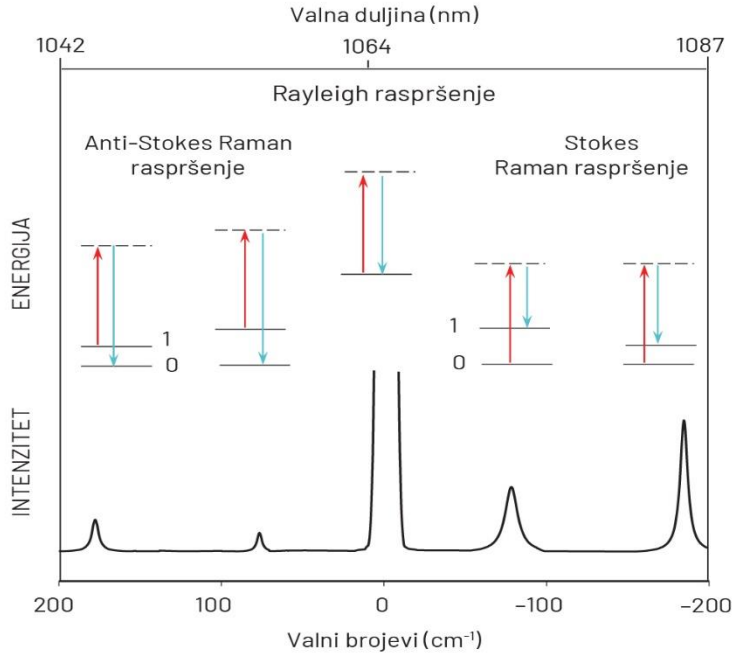
2.4.2. Ramanova spektroskopija

Ramanova spektroskopija temelji se na fenomenu neelastičnog (Ramanovog) raspršenja elektromagnetskog zračenja uslijed interakcije s uzorkom. Teoretski fenomen Ramanovog

raspršenja je prvi 1923. godine postavio Smekal (Smekal, 1928), a eksperimentalno su ga 1928. godine prvi put dokazali profesor Raman i njegov student Krishnah (Raman i Krishnan, 1928) u Indiji te Landsbergand i Mandelstam (Landberg i Mendelstam, 1928) u Rusiji. Međutim, otkriće ovog fenomena je pripisano samo profesoru Ramanu i njegovom studentu Krishnahu, te je za otkriće profesoru Ramanu dodijeljena Nobelova nagrada 1930. godine, a nova tehnika je po njemu dobila ime (Mitsutake i sur, 2019). Prvi komercijalni Ramanov spektrometar izrađen je tek 1953. godine, nakon razvoja monokromatora uz korištenje žive pri valnoj duljini od 435,8 nm kao izvora zračenja. Izumom lasera 60-tih godina prošlog stoljeća i njegovom primjenom kao primarnog izvora zračenja, riješen je veliki problem koji je predstavljala fluorescencija uzorka kod primjene lampi živinog luka, koje su uzrokovale pobudu elektronskih a ne samo vibracijskih stanja (Mitsutake i sur, 2019). U novije vrijeme implementacija NIR lasera zajedno sa Fourierovim transformacijama Raman spektroskopije (FT Raman) te cijeli niz tehnoloških unaprjeđenja, dovela je do znatnog napretka Raman tehnike.

Fenomen rasipanja svjetlosti može se klasično opisati EM zračenjem nastalim oscilirajućim dipolima induciranim u molekuli EM poljem upadnog zračenja. Prilikom raspršenja EM zračenja nije nužno da zračenje ima energiju koja odgovara energetskoj razlici. Interakcijom EM zračenja i molekula može se dobiti elastično i neelastično raspršenje. Da bi vibracija molekule bila aktivna u Ramanovom spektru, tijekom vibracije mora doći do polarizibilnosti. Inducirana polarizacija zasniva se na pomaku težišta negativnog naboja elektronske ovojnice i pozitivnog naboja atomskih jezgri. Molekula time dobiva inducirani dipolni moment, kojeg je veličina određena udaljenošću obaju težišta elektriciteta i efektivnim nabojem u težištima. Inducirani dipolni moment molekule nastaje kao rezultat polarizibilnosti molekule. Polarizibilnost je svojstvo molekule da se polarizira u prisutnosti električnog polja tj. polarizibilnost je deformacija elektronskog oblaka oko molekule vanjskim električnim poljem (Larkin, 2018). Upravo promjena polarizibilnosti molekule uzrokuje pojavu karakterističnih vrpci u Ramanovom spektru. Pri sobnoj temperaturi većina molekula je u osnovnom vibracijskom stanju. Interakcijom laserskog zračenja s molekulom dolazi do pobude u tzv. virtualna stanja, koja nisu stvarna vibracijska stanja već nastaju uslijed promjene polarizibilnosti molekule. Većina fotona ne mijenja svoju energiju nakon sudara s molekulom (elasični sudar) tj. ne dolazi do izmjene energije, već raspršeni fotoni imaju istu energiju kao i pobuđeni. Takvo zračenje je Rayleighovo raspršenje. Vrlo mali broj fotona (jedan od 10^6-10^8) će uslijed sudara s molekulama izmijeniti energiju s njima i to je primjer neelastičnog sudara. Raspršivanje u kojem upadni foton izmjenjuje energiju s molekulom je poznato kao Ramanovo raspršenje. Ukoliko uslijed sudara fotona i molekule foton pobudnog zračenja preda energiju

molekuli, raspršeni foton će imati manju frekvenciju od pobudnog fotona i to se naziva Stokesovo raspršenje. Ukoliko uslijed sudara molekula preda energiju fotonu, raspršeni će foton imati veću frekvenciju od pobudnog i to se onda naziva anti-Stokesovo raspršenje (Šašić, 2007). Raspršeni fotoni uključuju uglavnom dominantno Rayleighovo raspršenje uz vrlo malu količinu Ramanovog raspršenog svjetla (Larkin, 2018) (Slika 6).



Slika 6. Shematski prikaz Rayleigh raspršenja te Stokes i anti-Stokes Ramanovog raspršenja. Frekvencija pobude lasera predstavljena je strelicama (crveno) u smjeru odozdo prema gore i s puno većom energijom od molekulskih vibracija (pune linije). Frekvencija raspršenog fotona (plavo, strelice odozgo prema dolje) nepromijenjena je u Rayleighovom raspršenju, ali je niže ili više frekvencije u Ramanovom raspršenju. Isprekidane linije (crno) označavaju tzv. „virtualno stanje“ (prilagođeno iz Larkin, 2018).

2.5. Kemometrijske metode

Kemometrija se razvija kao sub-disciplina u kemiji kroz više od trideset godina. Razvoj instrumenata i procesa dovelo je do povećanja potrebe za naprednim matematičkim i statističkim metodama. Pojam „kemometri“ dolazi od švedskog znanstvenika Svante Wold 1971. godine (Wold, 1995). Kemometrija se može definirati kao kemijska disciplina koja koristi matematiku, statistiku i formalnu logiku za: (a) dizajniranje ili odabir optimalnih

eksperimentalnih podataka; (b) pružanje maksimalno relevantnih kemijskih informacija analizom kemijskih podataka; i (c) stjecanje znanja o kemijskim sustavima (Hopke, 2003; Luypaert i sur, 2007).

Primjenom kemometrije u analitičkoj kemiji može se doći do odgovora na naredna pitanja: kako iz izmjerenih podataka dobiti relevantne informacije; kako predstaviti i prikazati te informacije i kako integrirati takve informacije u podatke (Rodionova i Pomerantsev, 2006). Snažan razvoj kemometrije u kasnim 1970-ima povezan je sa primjenom računala, koji su postali dostupni za znanstvene i inženjerske svrhe (Rodionova i Pomerantsev, 2006).

Primjena kemometrije je doživjela veliki rast u 21. stoljeću razvojem različitih programa, koji su omogućili sasvim nove mogućnosti obrade dobivenih podataka, i danas kemometrija ima važnu ulogu u analitičkoj kemiji (Kumar i sur, 2014) i pruža iznimne mogućnosti u području analitičke kemije, kao discipline koja prati analitički tijek u svim fazama analize (Brereton i sur, 2018). Također je omogućen razvoj složenije opreme, koja omogućava veliki broj mjerenja, te primjena kompliciranih algoritama za analizu iznimno opsežnih skupova podataka. Međutim, pokazalo se da velika količina podataka ne mora nužno značiti da postoji dovoljno informacija (Rodionova i Pomerantsev, 2006). Analiza dobivenih podataka, koji se sastoje od brojnih varijabli izmjerenih iz više uzoraka, provodi se multivarijatnom analizom podataka. Cilj multivarijatne analize podataka je utvrditi sve varijacije u matrici podataka. Dakle, kemometrijski alati služe za pronalaženje veza između uzoraka i varijabli u danom skupu podataka i i pretvaranju u nove latentne varijable (Kumar i sur., 2014).

Kemometrija obuhvaća različite ciljeve. Jedan od ciljeva je predobrada eksperimentalnih podataka kako bi se poboljšala kvaliteta signala. Ostali ciljevi su konstrukcija modela za prepoznavanje obrazaca te kvantitativne metode određivanja (Pomerantsev, 2014). Za postizanje ovih ciljeva koriste se razne kemometrijske tehnike.

2.5.1. Predobrade eksperimentalnih podataka matematičkim metodama

Predobrada spektralnih podataka je sastavni dio kemometrijskog modeliranja. Matematičke metode predobrade su izuzetno korisne tehnike za početnu obradu tzv. sirovih spektralnih podataka. Cilj matematičkih predobrada spektara je smanjiti broj dimenzija, ukloniti nedostatke zbog raspršenja svjetlosti, ukloniti fizikalne pojave u spektru, slučajne šumove, odnosno eliminirati ili minimizirati varijabilnost koja nije povezana sa svojstvom od interesa, a sve kako bi se poboljšala daljnja multivarijantna regresija, klasifikacijski modeli ili eksploratorne

analize, tako da se odgovarajuće promjene mogu učinkovitije modelirati. Najčešće se koriste dvije kategorije metoda predobrade spektara: metode korekcije raspršenja (engl. *scatter correction*) i spektralne derivacije (Rinnan i sur., 2009). Raspršivanje se odražava kao pomak bazne linije spektra, koji može biti aditivan ili multiplikativan, ovisno o karakteristikama uzorka i fizičkoj interakciji uzorka sa svjetlošću (Martens i sur., 2003). Derivativne metode se prvenstveno koriste za preklapanje vrpci te uklanjanje velikih pomakla bazne linije. Također se koriste i za ispravljanje aditivnih i multiplikativnih varijacija bazne linije u spektru. Međutim, pri tome dolazi do pojačavanja šuma, što se može riješiti Savitzky-Golay algoritmom glačanja. Glačanje spektara uklanja slučajni šum iz spektralnih podataka i poboljšava vizualni aspekt spektra (Naes, 2002a). Pažljivim odabirom metode predtretmana spektara može se znatno poboljšati robustnost konačnog modela.

2.5.1.1. Metode derivacije

Derivacije pripadaju najčešćim predtretmanima spektralnih podataka i uglavnom se koriste kako bi se poboljšala rezolucija odnosno za rješavanje problema preklapanja vrpci te konstantnog i linearnog pomaka bazne linije među uzorcima. Najčešće se koriste prva i druga derivacija, a najveći nedostatak ove metode su pojačanje šuma i otežana spektralna interpretacija (Huang i sur, 2010).

Derivacija prvog reda (dy/dx) funkcionira po matematičkom principu koji pokazuje brzinu promjene zavisne varijable za svaku beskonačno malu promjenu u nezavisnoj varijabli. Derivacija drugog reda daje nagib prve derivacije, tako da je to zapravo brzina promjene nagiba izvornog spektra. Prva derivacija uklanja aditivni pomak bazne linije, što je vrlo korisno u NIR spektroskopiji. Vrpce kod prve derivacije su tamo gdje je maksimalni nagib izvornog spektra, i siječe os apscisu tamo gdje izvorni spektar ima vrpcu, pa je spektre nakon obrade prvom derivacijom teško interpretirati. Tendencija linearnog povećanja bazne linije spektara uklanja se primjenom druge derivacije, koja ima negativnu vrpcu tamo gdje originalni spektar ima vrpcu te je zbog toga lakši za interpretaciju. Primjena druge derivacija vrlo je učinkovita za rješavanje nagiba i zakrivljenosti bazne linije spektra.

Glačanje spektara Savitzky-Golay algoritmom uklapa spektar u polinom, a zatim uzima derivaciju tog polinoma. Korisna je kod vrlo oštrih apsorpcijskih vrpci s visokim šumom u spektru.

2.5.1.2. Korekcija višestrukog raspršenja

Korekcija višestrukog raspršenja (engl. *Multiplicative Scatter Correction*, MSC) također pripada u ponajviše korištene metode za korekciju spektara. Ovu su metodu uveli Martens i Jensen (1983). a doradili Geraldi i sur. (1985).

Pretpostavka na temelju koje se koristi MSC je da svi uzorci imaju isti koeficijent raspršenja za sve varijable. Korekcijski koeficijenti svakog spektra izračunavaju se regresijom na idealan spektar uzorka, odnosno, svaki spektar je što je moguće više prilagođen idealnom spektru uzorka (općenito prosječnom spektru), pomiče se i rotira kako bi se što više prilagodio odabranom referentnom, odnosno prosječnom spektru. Ostali spektri korigiraju se tako da imaju istu razinu raspršenja kao i prosječni spektar i to pomoću metode najmanjih kvadrata (Zeaiter i Rutledge, 2009). Osim toga, MSC metodom uklanja se i pomak bazne linije između spektara.

MSC se sastoji od dva koraka, kako slijedi:

1. Određivanje korekcijskih koeficijenata:

$$x_{org} = b_{ref,1}x_{ref} + b01 + e \tag{2.1}$$

2. Korigiranje izmjerenih spektara:

$$x_{corr} = \frac{x_{org} - b01}{b_{ref,1}} = x_{ref} + \frac{e}{b_{ref,1}} \tag{2.2}$$

gdje su:

x_{org} izvorni spektar uzorka;

x_{ref} je referentni spektar;

b0 i $b_{ref,1}$ su korekcijski koeficijenti, redom odsječak i nagib, izračunati linearnom regresijom u prvom koraku, a koriste se za ispravljanje svake vrijednosti spektra;

e je rezidual koji idealno predstavlja kemijsku informaciju u spektru;

x_{corr} je obrađeni tj. korigirani spektar;

1 je vektor čiji su svi elementi 1.

Glavni izazov MSC metode je potreba za određivanjem referentnog spektra (x_{ref}). Kao idealan x_{ref} spektar se najčešće koristi prosječni spektar svih snimljenih spektara (Rinnan i sur., 2009; Afseth i Kohler, 2012; Kohler, i sur., 2010).

Nedostatak MSC metode je da upravo zbog toga što se kao idealan spektar koristi srednji spektar, kemijske varijacije mogu utjecati na dobivene parametre. MSC je moguće proširiti u EMSC (engl. *extended MSC*), koja osim korekcije s obzirom na referentni spektar, može

uključivati dodatne korekcije valnih duljina ili poznatih spektralnih informacija (Martens et al., 2003).

2.5.1.3.Standardna normalna varijata

Standardna normalna varijata (engl. *Standard Normal Variate*, SNV) metoda korekcije spektara uklanja multiplikativne interferencije raspršenja zbog veličine čestica (Barnes i sur., 1989) te uklanja aditivna raspršenja bez promjene oblika izvornog spektra (Bocklitz i sur., 2011). Svaki spektar kod SNV metode standardiziran je vlastitom srednjom vrijednosti i standardnom devijacijom. SNV oduzima srednju vrijednost spektra i dijeli je sa standardnom devijacijom. Metoda uklanja varijacije nagiba za svaki pojedini spektar. Svaki spektar se transformira neovisno kao funkcija sukladno donjem izrazu:

$$x_{i,SNV} = \frac{x_i - \tilde{x}}{\sqrt{\sum(x_i - \tilde{x})^2/(n-1)}} \quad (2.3)$$

gdje su: $x_{i,\,snv}$ transformirani spektar; x_i izvorni spektar; $\tilde{x}$ je prosječna vrijednost varijable; n je broj varijabli u spektru.

Drugim rječima, SNV transformacija je centriranje redaka nakon čega slijedi skaliranje redaka (Pravdova i sur., 2001). Prednost ove metode je izostanak korištenja referentnog spektra, dok je nedostatak izostanak optimizacije u prvom koraku, zbog čega je u korigiranim podatcima moguća prisutnost bazne linije. SNV također korigira spektre s obzirom na raspršenje i pomak bazne linije.

2.5.1.4. Uklanjanje trenda

Uklanjanje trenda (engl. *de-trending*) je metoda spektralne korekcije, slična i često korištena zajedno sa SNV metodom, kojom se mogu smanjiti nelinearni trendovi postavljanjem polinomske jednadžbe za svaki spektar, koji se kasnije koristi za korekciju bazne linije. Uklanjanje trenda kod spektra izračunava se kao razlika između originalnog spektra i polinomske jednadžbe, koja opisuje novu baznu liniju (Zeaiter i Rutledge, 2009). Ovaj tip korekcije primjenjuje se na spektre kako bi se uklonile zakrivljenosti i pomaci bazne linije (Zeaiter i Rutledge, 2009). U kemometriji je uklanjanje trenda naziv metode za korekciju

spektra za nelinearno pomicanje osnovne linije pomoću polinomskih funkcija (Zeaiter i Rutledge, 2009).

2.5.1.5. Normalizacija

Svrha normalizacije je skalirati uzorke kako bi se dobili svi podaci na istoj skali. Ova se metoda primjenjuje kada se podaci prikupljaju metodom gdje je signal detektora funkcija mase uzorka (GC detektor) ili snage izvora (Ramanova spektroskopija) umjesto koncentracije uzorka. Nekoliko je metoda normalizacije dostupnih u kemometrijskim programskim paketima, kao što su: (1) normalizacija raspona koja je dobivena skaliranjem svih spektara prema njihovom rasponu; (2) normalizacija područja gdje je površina ispod spektra izjednačena (obično se koristi u kromatografiji); (3) srednja normalizacija, koja se dobiva dijeljenjem spektra sa srednjom vrijednošću za svaki spektar; i (4) maksimalna normalizacija, koja dijeli svaki red matrice podataka njegovom maksimalnom apsolutnom vrijednosti tj. maksimalnim apsolutnim spektrom za svaki spektar. Normalizacija koja se često koristi kod aplikacija prepoznavanja uzoraka je jedinična vektorska normalizacija, koja normalizira Xi podatke u jedinične vektore (Camo Analytics, 2014). Metode normalizacije koriste se kako bi se poboljšala točnost i prediktivne sposobnosti modela u multivarijantnoj analizi.

2.5.2. Eksploratorne analize podataka

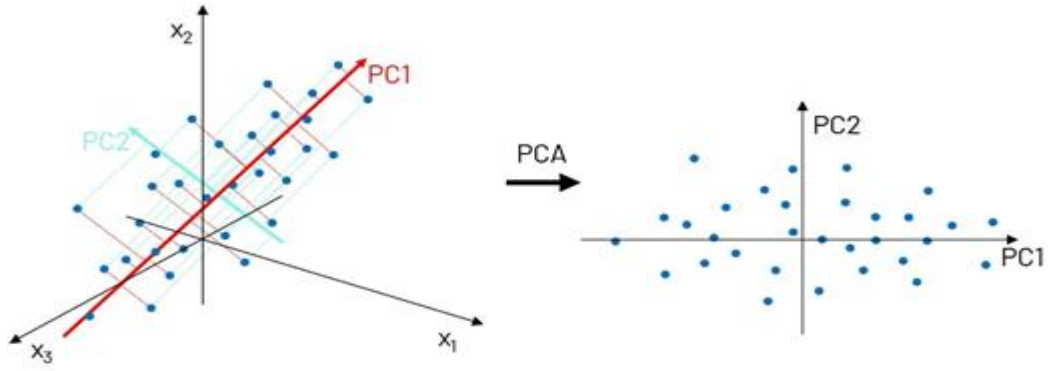
Nemoguće je raspravljati o kemometriji bez vizualizacije podataka (Geladi, 2003). Eksploratornom analizom podataka u kojoj se nastoje sažeti i što bolje vizualizirati podaci, započinje potpuna analiza podataka. Prije kalibracije modela, uvijek je poželjno procijeniti strukturu podataka i provesti eksploratornu analizu u cilju: (a) otkrivanja tzv. netipičnih uzoraka (engl. *outliera*); (b) prepoznavanja obrazaca u distribuciji uzoraka; i (c) procjene odnosa među varijablama i klasama (Ballabio i Consonni, 2013). Rezultati dobiveni eksploratornom analizom podataka koriste se za usmjeravanje daljnje analize podataka (npr. odabir varijabli relevantnih za klasifikaciju).

2.5.2.1.Analiza glavnih komponenata (PCA)

Analiza glavnih komponenata (engl. *Principal Component Analysis*, PCA) je osnovna metoda multivarijatne analize podataka, koja se često koristi za eksploratornu analizu podataka, identificiranje netipičnih uzoraka, smanjenje ranga (dimenzionalnosti) skupa podataka, grafičko grupiranje podataka, klasifikaciju i regresiju (Esbensen i Geladi, 2009). PCA je prvi prezentirao Pearson (1901). Pearson je proveo izračunavanja na skupovima od dvije do tri varijable. Opis izračunavanja kasnije je opisao Hotelling (1933). PCA je obično prvi korak u multivarijatnoj analizi podataka i često osnova za druge složenije tehnike prepoznavanja uzoraka među podacima. Koristi se za identificiranje grupa u podacima, komprimiranje skupa signala u interpretabilne varijable, koje se nazivaju glavne komponente (engl. *Principal Components*, PC), a zatim koristi ograničen broj značajnih PC-ova kao varijable za opisivanje uzoraka (Oliveri i Downey, 2012). PCA izdvaja relevantne informacije iz opsežnih podataka. Ono što se može smatrati informacijom ovisi o prirodi problema koji se rješava. Podaci ponekad sadrže informacije koje su potrebne, a u nekim slučajevima informacije mogu biti odsutne. Podaci (gotovo) uvijek sadrže neželjenu komponentu, koja se naziva šum. Priroda šuma varira, ali u mnogim slučajevima šum je dio podataka koji ne sadrži relevantne informacije. Koji dio podataka treba smatrati šumom, a koji dijelom informacije ovisi o ciljevima istrage i metodama koje se koriste za postizanje postavljenih ciljeva. Šum i suvišnost podataka se manifestiraju kroz korelacije između varijabli. Šum (pogreške u podatcima) dovodi do nesustavnih, slučajnih odnosa između varijabli (Pomerantsev, 2014).

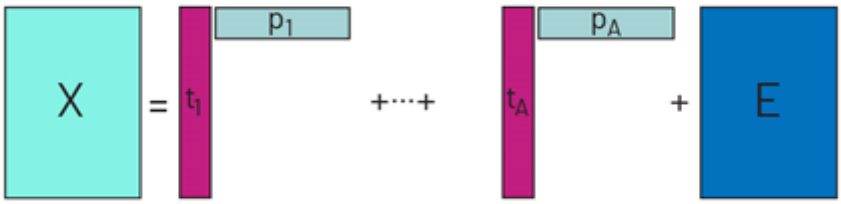
Centralna ideja PCA je redukcija dimenzionalnosti skupa podataka (izvornih informacija), koji se sastoji od velikog broja međusobno povezanih varijabli, a da se pri tome zadrži što je moguće više varijacija prisutnih u skupu podataka. PCA transformira skup podataka s koreliranim varijablama u skup nekoreliranih PC, koji se dobivaju kao linearna kombinacija početnih varijabli. PC se izračunavaju tako da se maksimalan udio varijance objašnjava prvom glavnom komponentom, a postupno svaka slijedeća komponenta objašnjava manji udio varijance (Brereton i sur., 2017;Pomerantsev, 2014). Kao rezultat opisanoga postupka podaci su predstavljeni u novom prostoru, čija je dimenzija mnogo niža od izvornog broja varijabli (Slika 7.) Uporaba PCA bavi se vizualizacijom latentnih struktura podataka pomoću grafičkih prikaza odnosno trendova podataka u dimenzionalnom prostoru, čija interpretacija otvara dublje razumijevanje od onoga što je moguće kada se promatraju samo pojedine varijable, jer PCA omogućuje interpretacije na temelju svih varijabli istovremeno (Esbensen i Geladi, 2009). Grafički prikaz koristan je za vizualiziranje dimenzionalnog prostora i isticanje informacija

koje bi se mogle upotrijebiti za utvrđivanje međugrupnih i unutargrupnih razlika (Wold i sur., 1987). Korelacijska matrica je početna točka PCA i ona daje linearne funkcije varijabli, a koje imaju svojstvo da nisu u korelaciji jedna s drugom i da su poredane prema količini ukupne varijacije koju obuhvaćaju (Bartholomew, 2010).



Slika 7. Princip PCA (prilagođeno iz Abdi i Williams, 2010). x_1; x_2; x_3 -varijable; PC1; PC2-glavne komponente.

Problem kolinearnosti, koji se javlja kada je korelacija između varijabli velika, rješava se tako da se podaci reduciraju na manji skup od samo nekoliko latentnih varijabli, koje najbolje opisuju izvorne podatke tj. objašnjavaju maksimalnu varijancu u podatcima (Wold i sur., 1987). Redukcija podataka se postiže dekompozicijom izvorne matrice X (I x J), koja se razlaže na produkt dvije matrice. Izvorna matrica X zamijeni se sa dvije matrice T i P, čija je zajednička dimenzija A manja od broja varijabli (stupci) J u matrici X (jednadžba 2.4.). Očuvana je druga dimenzija matrice T, odnosno broj uzoraka (redovi) I (Slika 8.). Ukoliko se razlaganje matrice X pravilno provede, odnosno dimenzionalnost A se pravilno odredi, matrica T će sadržavati iste podatke koji su sadržani u inicijalnoj matrici X, ali je matrica T manja i stoga manje složena od matrice X (Pomerantsev, 2014).



Slika 8. Razlaganje PC (prilagođeno iz Pomerantsev, 2014). X – inicijalna matrica; t_1 - t_A - projekcije uzoraka u novom koordinatnom sustavu; p_1 - p_A; - projekcija varijabli u novom koordinatnom sustavu; - ; E – matrica reziduala.

$$X = TP^t + E \tag{2.4}$$

gdje su:

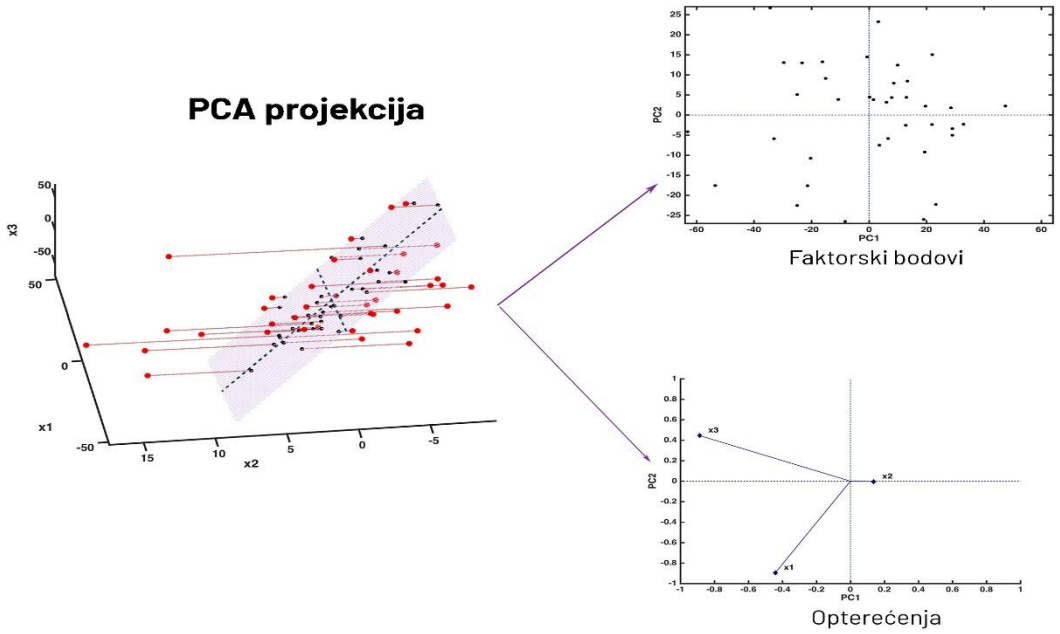
T - matrica faktorskih bodova dimenzija (IxA)

P - matrica opterećenja dimenzija (JxA);

E - matrica reziduala dimenzija (IxJ). (Pomerantsev, 2014);

U multivarijatnoj analizi podataka, ovu matricu treba učiniti malom. Razlog tome je što sadrži uglavnom šum mjerenja i uzorkovanja (Esbensen i Geladi, 2009). Nove se varijable nazivaju PC, od toga i naziv metode PCA.

Važno svojstvo PCA je ortogonalnost (neovisnost) PC, što znači da se matrica T ne obnavlja kada se broj komponenata poveća. Matrica T se samo proširi za još jedan stupac, koji odgovara novom smjeru PC. Isto se događa s matricom opterećenja P. A je ključna karakteristika koja definira složenost podataka (Pomerantsev, 2014).



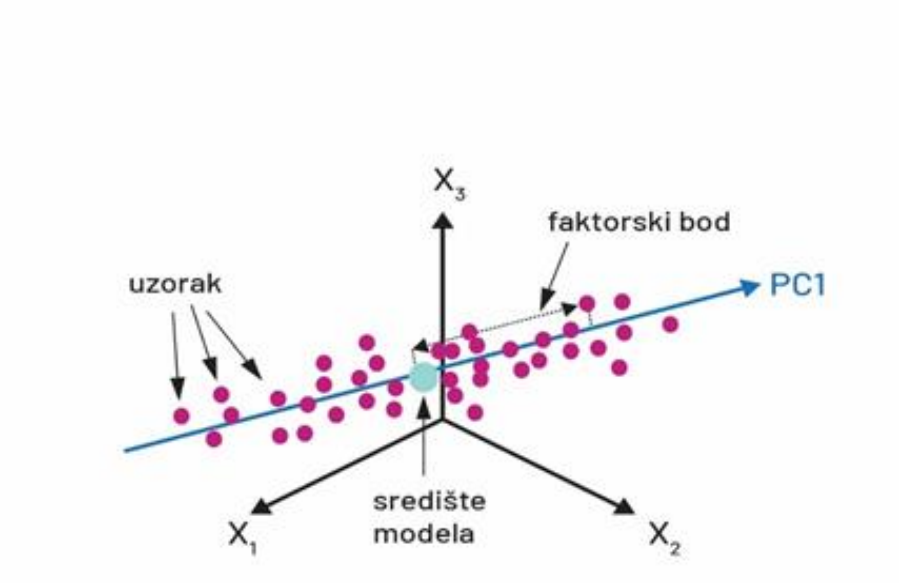
Slika 9. Shematski prikaz PCA (prilagođeno iz Biancolillo i Marini, 2018).

Na slici 9. uzorci predstavljeni u trodimenzionalnom prostoru, projiciraju se u nisko-dimenzionalni podprostor (označeni svijetlocrvenom bojom) obuhvaćeni s prve dvije glavne komponente. Inspekcija skupa podataka može se provesti promatranjem distribucije uzoraka na informativni PC podprostor (grafikon faktorskih bodova), a interpretacija se može načiniti procjenom relativnog doprinosa eksperimentalne varijable u definiranju glavnih komponenata. (grafikon opterećenja; Biancolillo i Marini, 2018). Četiri su dijela PCA modela: podaci, faktorski bodovi (engl. *scores*), opterećenja (engl. *loadings*) i reziduali (Bro i Smilde, 2014).

2.5.2.1.1. Faktorski bodovi

Matrica faktorskih bodova T sadrži projekcije uzoraka (J-dimenzionalni redni vektori x_1, …, x_I) u A dimenzionalnom podprostoru PC. Redovi matrice T (t_1, ..., t_I) su koordinate uzoraka u novom koordinatnom sustavu. Stupci t_1, ..., t_A matrice T su ortogonalni i svaki stupac t_A sadrži projekcije svih uzoraka na a-tu os novih koordinata (Pomerantsev, 2014). Faktorski bodovi predstavljaju informacije korisne za razumijevanje strukture podataka (Pomerantsev, 2014), a grafički prikaz faktorskih bodova (engl. *score plot*) naziva se i mapa uzoraka. Mapa uzoraka

prikazuje povezanost među uzorcima koji se istražuju (Brereton i sur., 2017) i glavni je alat PC analize (Slika 10.; Pomerantsev, 2014).

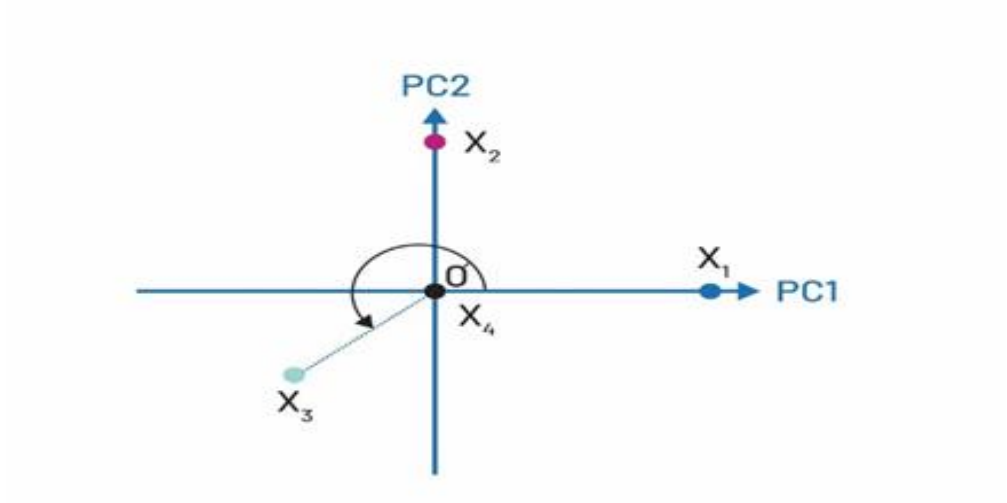


Slika 10. Shematski prikaz značenja faktorskog boda (preilagođeno iz Camo Analytics, 2014).

2.5.2.1.2. Opterećenja

Matrica opterećenja P je prijelazna matrica iz izvornog J-dimenzionalnog prostora varijabli x_1, …, x_J u A-dimenzionalni prostor PC. Svaki red matrice P sadrži koeficijente koji se odnose na varijable t i x. Tako je npr. A-ti red projekcija svih varijabli x_1, …, x_J na a-tu PC os. Svaki stupac matrice P je projekcija povezane varijable x_J na novoformirani koordinatni prostor (Pomerantsev, 2014). Grafikon opterećenja također se naziva i mapa varijabli i ova mapa prikazuje utjecaj i međusobne veze varijabli u skupu podataka (Brereton i sur., 2017), a koristi se za proučavanje utjecaja varijabli. Analiza grafikona opterećenja slična je grafičkom prikazu faktorskih bodova (Slika 11.). Analizom grafikona opterećenja moguće je uočiti koje varijable koreliraju a koje su neovisne (Pomerantsev, 2014). U geometrijskom smislu, opterećenja predstavljaju kosinus kuta između varijable i trenutnog PC. Što je kut manji (tj. veća veza između varijable i PC), to je opterećenje veće. Ove vrijednosti mogu varirati između -1 i +1.Opterećenja se koriste kako bi se dobio uvid u korelaciju među varijablama. Ako varijable imaju velika optrećenja na istom PC (njihov kut je malen), ove su dvije varijable u velikoj

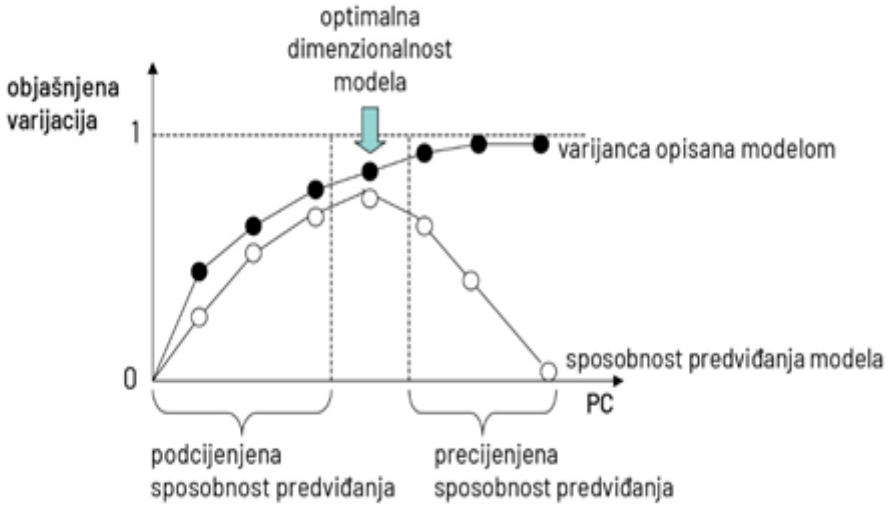
korelaciji. Ako oba opterećenja imaju isti predznak (- ili +), korelacija je pozitivna (kada se jedna varijabla povećava, povećava se i druga). U suprotnom je korelacija dviju varijabli negativna (kada jedna varijabla raste, druga se smanjuje; Slika 11.) (Camo Analytics, 2014). PC grade vezu između uzoraka i varijabli pomoću faktorskih bodova i opterećenja (Camo Analytics, 2014).



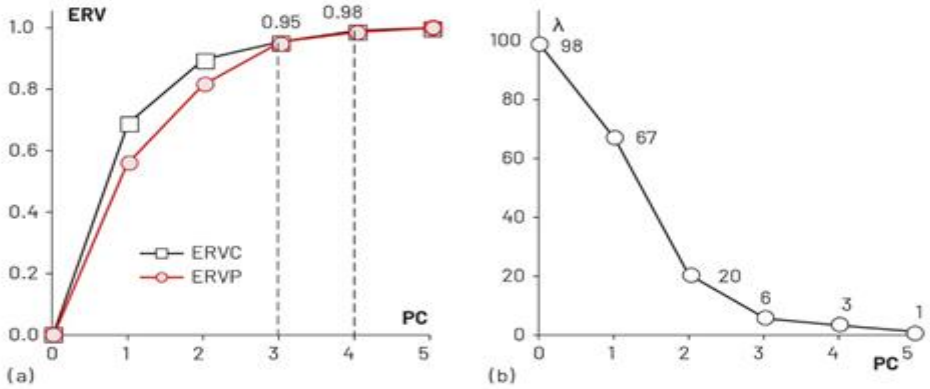
Slika 11. Shematski prikaz značenja opterećenja u vidu korelacije varijabli i glavnih komponenti (Camo Analytics, 2014). $x_1 – x_4$ su varijable.

2.5.2.1.3. Određivanje broja glavnih komponenti (PC)

PCA je iterativni postupak kojim se nove komponente redom dodaju jedna po jedna. Važno je odlučiti kada ovaj postupak treba prekinuti, odnosno odrediti odgovarajući broj PC. Ako je broj PC premali, opis podataka će biti nepotpun. S druge strane, pretjeran broj PC rezultira precijenjenom sposobnosti predviđanja modela (engl. *overfitting*) tj. modeliranjem šuma, a ne bitnih informacija (Slika 12.). Obično se broj PC odabire uz korištenje grafičkog prikaza (Pomerantsev, 2014) na kojem je objašnjena rezidualna varijanca (engl. *Explained Residual Variance*), ERV), koja je prikazana kao funkcija PC brojeva.



Slika 12. Utjecaj dimenzionalnosti modela na sposobnost prilagođavanja i predviđanja (prilagođeno iz Wiberg, 2004).



Slika 13. Odabir broja PC (prilagođeno iz Pomerantsev, 2014). ERV- objašnjena rezidualna varijanca (engl. *Explained Residual Variance*); ERVC - objašnjena kalibracijska rezidualna varijanca; ERVP - objašnjena validacijska rezidualna varijanca λ - svojstvene vrijednosti (engl. *eigenvalues*).

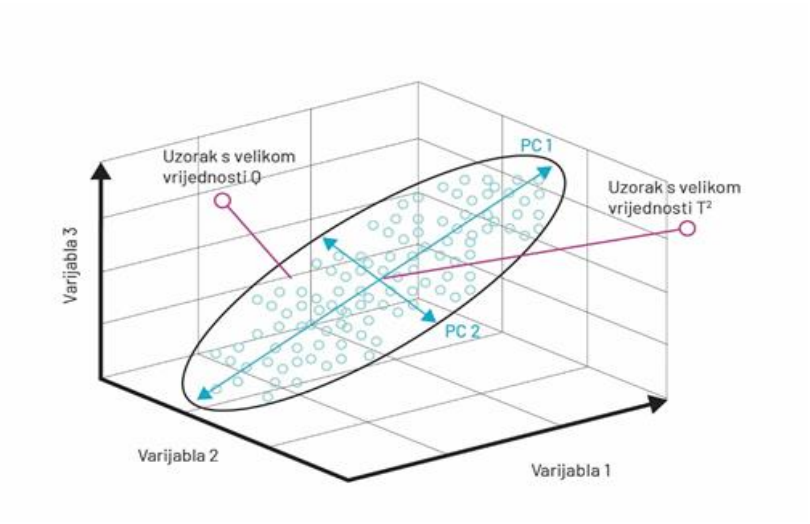
Na Slici 13. kao primjer odabira broja PC prikazana je objašnjena rezidualna varijanca u kalibraciji (engl. *Explained Residual Variance in training*, ERVC) i u validaciji (engl. *Explained Residual Variance in validation*, ERVP). Iz ove se slike može zapaziti: (a) da tri PC obuhvaćaju 95 % inicijalne varijacije, dok četiri PC-a obuhvaćaju 98 % inicijalne varijacije, a

konačno odlučivanje o broju PC donosi se tek nakon analize podataka; (b) oštra promjenu zakrivljenosti linije (prekid) pri PC 3, pa je stoga optimalan broj PC 3 ili 4. λ-svojstvene vrijednosti, karakterizira važnost svake komponente (matrica faktorskih bodova T).

2.5.2.1.4. Netipični i ekstremni uzorci

Pogreške i neočekivane pojave u stvarnom su svijetu neizbježne. To vrijedi i za primjenu kemometrijskih tehnika. Uvijek postoje određeni uzorci koji se iz nekog razloga razlikuju u odnosu na ostatak skupa podataka (Naes i Isaksson, 2002) Netipični uzorci (engl. *outliers*) su uzorci koji pokazuju netipična svojstva u usporedbi sa ostalim objektima u skupu podataka. Bez obzira na uzrok pojavljivanja, netipičnim uzorcima treba pristupati sa posebnom pažnjom kako analiza podataka ne bi dovela do pogrešne klasifikacije uzoraka. Također, važno je netipične uzorke ne miješati sa ekstremnim uzorcima, koji su uvijek prisutni u podacima (Pomerantsev i Rodionova, 2014). Inicijalno, prisutnost netipičnih uzoraka u modelu se može procijeniti pomoću nekoliko indeksa, poput utjecajnih vrijednosti, Q reziduala i Hotelling T^2 statistike (Ballabio i Consonni, 2013; Bro i Smilde, 2014).

Također relevantni alati za analizu i razumjevanje obrazaca podataka su dijagram faktorskih bodova (engl. *score plot*) i dijagram opterećenja (engl. *loading plot*). Dijagram faktorskih bodova je grafička vizualizacija podatkovnih točaka u reduciranom prostoru definiranom sa odabranim PC (Slika 14.). Da bismo razumjeli koliko informacija se prikazuje, treba uzeti u obzir objašnjenu varijancu (Ballabio i Consonni, 2013).



Slika 14. PCA model sa označenim Q i T^2 netipičnim uzorcima (prilagođeno iz Mujica i sur., 2010).

Uzorci s velikim utjecajem, odnosno uzorci koji imaju visoku utjecajnu vrijednost (engl. *leverage*) su mjera utjecaja uzorka u modelu i mogu znatno utjecati na kvalitetu PCA modela. Utjecajna vrijednost opisuje koliko daleko je lociran uzorak u prostoru faktorskih bodova. Uzorci visoke utjecajnosti, odnosno utjecajne vrijednosti predstavljaju potencijalne netipične uzorke i treba ih promatrati s posebnom pažnjom (Marini, 2013). Uzorci visoke utjecajne vrijednosti također su usko povezani sa Mahalanobisovom udaljenosti (Naes i Isaksson, 2002). Za interpretaciju i karakterizaciju svakog objekta važne su dvije statistike. Jedna je Hotelling statistika H (često označena kao T^2), koja je jednaka kvadratu Mahalanobisove udaljenosti od središta modela do projekcije uzorka unutar PCA podprostora. To se naziva bodovna udaljenost (engl. *score distance*, SD). Druga statistika je ortogonalna udaljenost (engl. *orthogonal distance*, OD), Q, također poznata i kao kvadratna pogreška predviđanja (engl. *squared prediction error*, SPE), što je kvadrat Euklidske udaljenosti od uzorka do PCA podprostora (Pomerantsev i Rodionova, 2020).

Q statistika pokazuje koliko se dobro novi uzorak uklapa u PCA model formiran na prethodnim mjernim podacima. To je mjera razlike (rezidualne) između uzorka i njegove projekcije na PC zadržane u modelu (Slišković i sur., 2012; Qin, 2003).

Q statistika se koristi za procjenu koliko je svaki uzorak u skladu s modelom. To je mjera razlike između izvornih podataka i podataka rekonstruiranih na temelju kalibriranog modela. Q vrijednosti pridružene su svakom uzorku i velike Q vrijednosti označavaju uzorke koji imaju velike reziduale izvan modela. Hotelling T^2 statistika temelji se na zbroju normaliziranih kvadratnih faktorskih bodova, što je mjera varijacije u svakom uzorku unutar modela. Slično tome, utjecajni uzorak je mjera udaljenosti uzorka od središta modela (Slika 14.; Ballabio i Consonni, 2013a; Wise i sur., 1997).

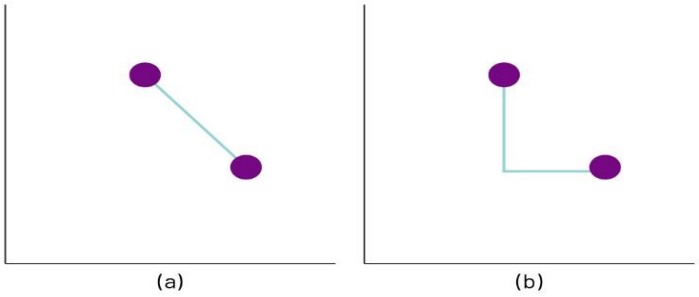
Obje velike vrijednosti Hotelling T^2 i utjecajne vrijednosti ukazuju na uzorak koji ima veliki utjecaj na model. Također se može uspostaviti granica pouzdanosti (engl. *confidence limit*) za Q, T^2 i utjecajne uzorke. Hotelling T^2 i utjecajni uzorci imaju različite skale, ali su uvijek u korelaciji. Važno je izbjegavati prisutnost uzoraka s prevelikim utjecajem, odnosno značajno velikim vrijednostima T^2. Općenito, ako se identificiraju netipični uzorci, treba ih analizirati kako bi se pojasnio razlog zašto se razlikuju od ostatka podataka i nakon ove procjene važno je odlučiti hoće li ih se zadržati ili ukloniti iz analize (Ballabio i Consonni, 2013).

2.5.2.2. Klasterska analiza

Klasterska analiza (engl. *Cluster Analysis, CA*) ili analiza grupiranja koristi se za određivanje i prepoznavanje temeljnih obrazaca i strukture naizgled besmislenih podataka. Osnovni cilj je prirodno grupiranje podataka sličnih karakteristika (Adams, 2004; Of i sur., 2011), odnosno cilj klasteriranja je odrediti unutarnje grupiranje u skupu neoznačenih podataka (Abonyi i Feil, 2007). Kod klasterske analize grupna pripadnost objekata nije poznata, kao ni konačan broj grupa (Einax i sur., 1998).

Prvi korak u klasterskoj analizi je utvrđivanje sličnosti među objektima. Četiri su najpopularnija načina određivanja međusobne sličnosti (Brereton, 2018):

- **Koeficijent korelacije među uzorcima**. Koeficijent korelacije 1 implicira da uzorci imaju identične karakteristike. Što je koeficijent korelacije negativniji, to su objekti manje slični. Što je koeficijent korelacije veći, to su objekti sličniji.
- **Euklidska udaljenost**. Što je manja Euklidska udaljenost, to su uzorci sličniji. Ova udaljenost djeluje suprotno koeficijentu korelacije i mjera je različitosti. Dok se koeficijent korelacije kreće uvijek od -1 do +1, to ne vrijedi za Euklidsku udaljenost, koja nema ograničenja, iako je uvijek pozitivan broj (Slika 15.a).
- **Manhattan udaljenost**. Manhattan udaljenost je nešto drugačije definirana od Euklidske udaljenosti i uvijek je veća od Euklidske udaljenosti (Slika 15.b).
- **Mahalanobisova udaljenost**. Mahalanobisova udaljenost je metoda popularna kod kemičara, iako je površno slična Euklidskoj, uzima u obzir korelaciju među varijablama.



Slika 15. Shematski prikaz (a) Euklidske udaljenosti i (b) Manhattan udaljenosti između objekata.

Slijedeći korak u klasterskoj analizi je povezivanje objekata. Najčešći pristup u povezivanju objekata je aglomerativno grupiranje, u kojem se pojedinačni objekti postupno međusobno

povezuju u skupine (Brereton, 2018). Pri tome se često koristi hijerarhijsko aglomerativno grupiranje. Dvije su vrste hijerarhijskih metoda grupiranja - aglomerativne i dijeleće (engl. *divisive*). Aglomerativna hijerarhijska metoda započinje s objektom kao vlastitim klasterom. Zatim sukcesivno spaja najsličnije klastere zajedno, dok čitav skup podataka ne postane jedna grupa (Wold i sur, 2014).

Prvi korak aglomerativne hijerarhijske metode je iz sirovih podataka pronaći dva najsličnija objekta (najbliža), a zatim se formira skupina od ova dva najbliža objekta (Abonyi i Feil, 2007). Važno je odlučiti kako predstaviti ovo novo grupiranje. Glavni slijedeći zadatak je ponovni izračun brojčane vrijednosti sličnosti između nove skupine i preostalih objekata. Postoje različiti načini za to kako je pobrojeno i opisano ovdje ispod (Brereton, 2018; Abonyi i Feil, 2007; Murtagh i Contreras, 2012).

Postupak jednostrukog povezivanja *(*engl. *single linkage)* ili tzv. najbližeg susjeda (engl. *nearest neighbour*). Udaljenost između dva klastera najmanja je udaljenost između svih parova uzoraka iz dva klastera, pri čemi je jedan uzorak iz prvog, a drugi iz drugog klastera. Sličnost nove skupine sa svim ostalim skupinama data je najvećom sličnosti bilo kojeg od izvornih objekata međusobno.

Postupak potpunog povezivanja (engl. *complete linkage*) ili tzv. najdalji susjed (engl. *farthest neighbour*) Ovaj je postupak suprotno od najbližeg susjeda i ovdje se koristi najmanja sličnost. Metoda najudaljenijeg susjeda odnosi se samo na izračun mjera sličnosti nakon što se formiraju nove skupine. Dvije skupine (ili objekti) s najvećom sličnosti i dalje su uvijek prvi pridruženi.

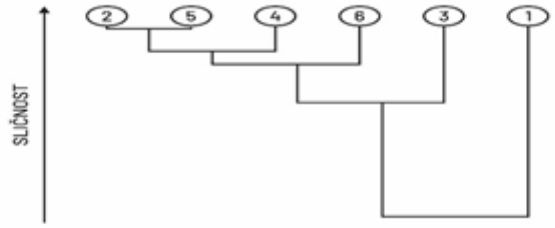
U oba navedena slučaja, dva klastera se spajaju kako bi formirali veći klaster na temelju kriterija minimalne udaljenosti (Abonyi i Feil, 2007). Naime, uočeno je da algoritam potpunog povezivanja formira korisnije hijerarhije, puno je pragmatičniji u mnogim aplikacijama od postupka jednostrukog povezivanja (Abonyi i Feil, 2007).

Prosječna povezanost (engl. *average linkage*). Prosječna povezanost spaja skupine na temelju prosječne udaljenosti svih objekata u jednoj grupi do svih objekata u drugoj grupi (Ferreira i Hitchcock, 2009).

Ward metoda (Ward, 1963). Wardova metoda, koja se ponekad naziva i metodom minimalne varijance, nadmašuje ostale hijerarhijske metode grupiranja. Ova se metoda temelji na pojmovima kvadratne pogreške populariziranim u analizi varijance ili drugim statističkim postupcima (Dubes, 1988). Ward je jedina među aglomerativnim metodama klasteriranja koja se temelji na klasičnom kriteriju zbroja kvadrata, formirajući skupine koje minimiziraju disperziju unutar grupe pri svakoj binarnoj fuziji. Uz to, Wardova metoda zanimljiva je jer klastere traži u multivarijantnom Euklidskom prostoru, koji je također referentni prostor u

višavarijantnim metodama, a posebno u PCA metodi (Wold i sur., 2014). Iako je Wardova metoda slična metodama povezivanja po tome što započinje s N klastera, od kojih svaki sadrži jedan objekt, razlikuje se od navedenih metoda po tome što ne koristi udaljenosti klastera za grupiranje objekata. Umjesto toga, ovom se metodom izračunava ukupni zbroj kvadrata unutar klastera kako bi se utvrdile slijedeće dvije skupine spojene u svakom koraku algoritma (Wold i sur., 2014). Wardova metoda daje jasno strukturirane i relativno stabilne klastere u širokom rasponu sličnosti (Einax i sur., 1998).

Slijedeći koraci hijerarhijskog aglomerativnog grupiranja sastoje se od nastavka grupiranja podataka, sve dok se svi objekti ne pridruže jednoj velikoj grupi. U svakom se koraku identificira najsličniji par objekata ili klastera, a zatim se kombiniraju u jedan novi klaster, sve dok se svi objekti ne spoje. U nekim slučajevima može se formirati nekoliko klastera, iako se u konačnici obično formira jedna velika skupina. Normalno je tada odrediti kojom mjerom sličnosti se svaki objekt pridružio većoj skupini i koji objekti međusobno najviše nalikuju (Brereton, 2018; Abonyi i Feil, 2007). Uz jednostavnost i činjenicu da se temelji na prirodnom kriteriju grupiranja, hijerarhijsko aglomerativno grupiranje često dolazi s popularnim grafičkim prikazom, koji se naziva dendrogram, a koristi se kao podrška za odabir modela (izbor broja klastera) i interpretaciju rezultata (Randriamihamison i sur., 2020; Brereton, 2018). Dendogram pripada u najinformativnije i najjednostavnije metode predstavljanja klaster analize (Bratchell, 1992). Objekti su organizirani u nizu, u skladu sa sličnostima (Slika 16.): okomita os predstavlja mjeru sličnosti kojom se svaki uzastopni objekt pridružuje grupi (Brereton, 2018).



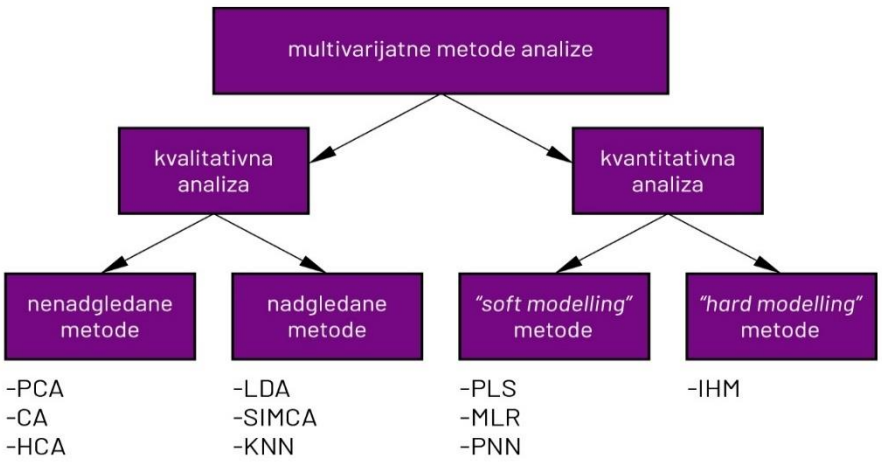
Slika 16. Dendogram hijerarhijskog klasteriranja. 1 - 5 su objekti.

2.5.3. Multivarijatne klasifikacijske metode

Multivarijatne klasifikacijske metode su kemometrijske tehnike čiji je cilj pronalaženje matematičkih modela koji mogu sortirati objekte u skupine (klase) u skladu sa numeričkim vrijednostima varijabli (obilježja), koje karakteriziraju svojstva tih objekata, a na temelju niza mjerenja. Jednom kada se klasifikacijski model formira, može se predvidjeti pripadnost nepoznatih uzoraka jednoj od definiranih klasa. Stoga se klasifikacijske tehnike bave kvalitativnim odgovorima tj. utvrđuju matematičke veze između skupa deskriptivnih varijabli (npr. kemijska mjerenja) i kvalitativnih varijabli tj. pripadnost definiranoj kategoriji (Ballabio i Consonni, 2013; Pomerantsev, 2014).

Početni podaci za klasifikaciju sadržani su u matrici X, gdje svaki redak predstavlja objekt, a svaki stupac odgovara varijabli. Broj objekata (redovi X) označen je slovom I, broj varijabli (stupci X) slovom J, a broj klasa označen je s K. Pojam klasifikacije primjenjuje se ne samo na sam postupak dodjele već i na njegov rezultat. Metoda (algoritam), koja provodi klasifikaciju, naziva se klasifikator (Pomerantsev, 2014).

Metode klasifikacije mogu se podijeliti u dvije vrste (Slika 17.) - (1) nadgledani (engl. *supervised*) pristupi pokušavaju podijeliti objekte u skupine prema njihovim karakteristikama pomoću trening seta, odnosno objekata koji su označeni u unaprijed definiranim kategorijama, i (2) nenadgledani (engl. *unsupervised*) pristupi koji pokušavaju podijeliti podatkovni prostor u grupe bez unaprijed definiranog trening seta (Brereton, 2015). Sinonim za klasifikaciju može se smatrati pojam prepoznavanje uzorka (engl. *pattern recognition*).



Slika 17. Multivarijatne metode analize.

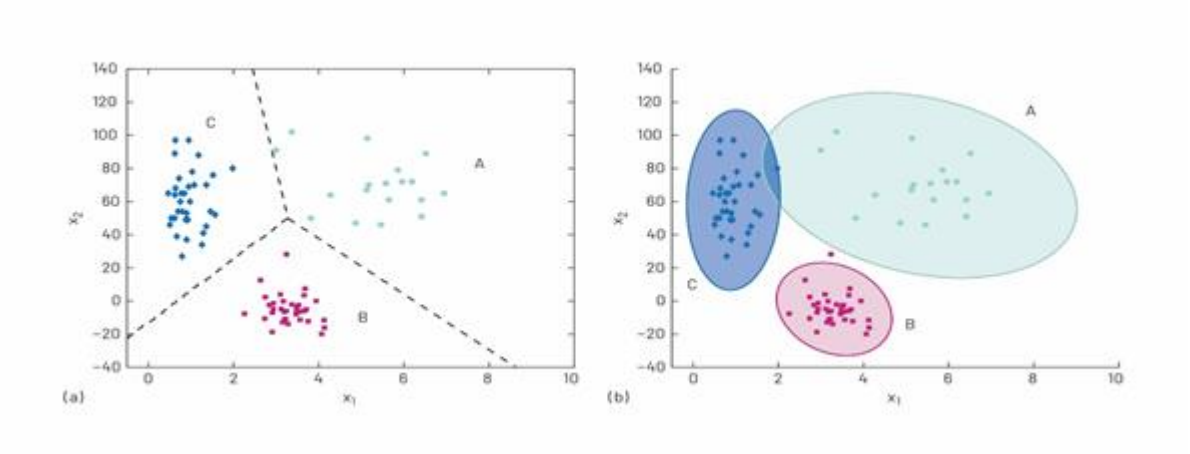
2.5.3.1. Metode klasnog modeliranja i diskriminantne analize

Kvalitativno modeliranje podataka je temeljna grana prepoznavanja uzoraka među podacima i obuhvaća dva glavna pristupa (Slika 18.) - diskriminantne i metode klasnog modeliranja (Oliveri, 2017).

Diskriminantne metode i metode klasnog modeliranja su nadzirane (engl. *supervised*) klasifikacijske tehnike, što znači da prije bilo kakvog modeliranja treba prikupiti skup reprezentativnih uzoraka. Ovaj skup bi trebao uključivati uzorke svake klase uključene u klasifikaciju, a čija je pripadnost klasi neosporna. Trebao bi također postojati dodatni podskup, nazvan test set, koji je sličan trening setu, ali je manji od ovoga potonjeg. Test set se koristi za optimizaciju i provjeru valjanosti modela. U slučaju kada taj skup nije dostupan, može se primijeniti tehnika unakrsne validacije (engl. *cross validation*, CV; Esbensen i Geladi, 2010; Rodionova i Pomerantsev, 2020). Klasa (ili kategorija) definira se kao skupina uzoraka koja imaju jedno ili više zajedničkih svojstava. Ta se svojstva obično mogu opisati matematičkim varijablama, pa je stoga moguće konstatirati da uzorke, koji čine klasu, karakteriziraju jednake vrijednosti diskretnih varijabli ili slične vrijednosti (unutar određenog raspona) kontinuiranih varijabli. Ako su te varijable, koje definiraju pripadnost klasi, lako mjerljive za svaki pojedini uzorak, dodjeljivanje novih uzoraka klasi je izravan i automatski zadatak. Suprotno tome, ako se takve varijable ne mogu izmjeriti na jednostavan način, pripadnost klasi se ne može izravno odrediti. Kako bi se riješila ova situacija, metode klasifikacije uspostavljaju i koriste matematičke odnose između ostalih varijabli, koje je lako izmjeriti, i klasne pripadnosti. To je moguće u slučaju kada te varijable sadrže korisne informacije i ako je na raspolaganju određeni broj uzoraka određene klasne pripadnosti za formiranje klasifikacijskog modela (Oliveri, 2017). Bilo koje druge klase objekata, koji nisu pripadnici ciljne klase, su alternativne klase ili uopće ne pripadaju nijednoj određenoj klasi. Kod klasnog modeliranja za razliku od diskriminantne analize, takvi objekti se ne koriste za modeliranje ciljne klase. Ciljna klasa označena je svojstvima svojih reprezentativnih članova. Ta su svojstva, koja se nazivaju i „otisci prstiju“, multivarijatni analitički signali dobiveni spektroskopijom, kromatografijom ili drugim analitičkim tehnikama (Rodionova i sur., 2016).

Prilikom odabira strategije klasifikacije, važno je definirati da li određeni problem, koji se istražuje, dopušta višeklasni ili samo jednoklasni izbor. Diskriminantne metode koriste se samo za rješavanje višeklasnih situacija, dok se klasno modeliranje može koristiti za jedoklasne i višeklasne probleme (Oliveri, 2017).

Klasno modeliranje provjerava usklađenost sa specifikacijom, definirajući zatvoreni multivarijatni prostor klase pri unaprijed određenoj razini pouzdanosti i to za autentične uzorke klase koja se istražuje. Tako formirani modeli imaju prednost opisivanja ciljnih uzoraka bez distribucije ne ciljnih uzoraka u trening setu. Suprotno tome, diskriminantne metode traže graničnik između dviju ili više klasa, koristeći doprinos svih razmatranih klasa. To znači da sve klase moraju biti točno definirane i da uključeni uzorci moraju biti reprezentativni za svaku klasu, jer imaju presudan utjecaj na odluku o pripadnosti uzorka odgovarajućoj klasi.(Oliveri, 2017; Oliveri i Downey, 2012).



Slika 18. Diskriminacijska klasifikacija (a) i klasno modeliranje (b) (prilagođeno iz De Luca i sur., 2018). A, B i C označavaju različite klase uzoraka.

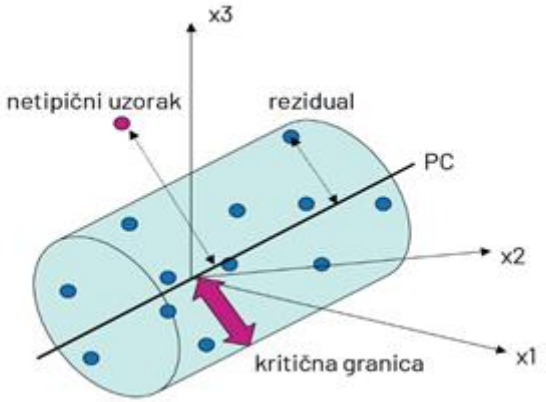
U slučaju jednoklasne klasifikacije, da bi se izgradio klasifikacijski model, potrebni su samo uzorci ciljne klase. Dodatni uzorci koji ne pripadaju ciljnoj klasi pomažu kod validacije modela. Izbor alternativnih klasa ne bi trebao biti slučajan, već zasnovan na tzv. konceptu najbližeg roda (engl. *Nearest of Kin*, NoK). NoK klase su najbliži analozi ciljne klase, a to su npr. uzorci proizvedeni na istom mjestu, sličnom tehnologijom i pravilima kontrole kvalitete (Rodionova i sur., 2019).

Meko neovisno modeliranje analogne klase (engl. *Soft Independent Modelling of Class Analogy*, SIMCA) je primjer jednoklasnog klasifikatora, koji daje opis ciljne klase objekata, a zatim upućuje je li je novi objekt sličan ovoj klasi ili ne. Suprotno tome, primjer diskriminantne metode, Diskriminantna analiza parcijalnih najmanjih kvadrata (engl. *Partial least squares discriminant analysis*, PLS-DA) daje opis nekoliko skupova objekata, koji predstavljaju predefinirane klase, a zatim određuje članstvo objekta u jednoj od tih klasa. Stoga, nije

dosljedno uspoređivati metode koje imaju različite ciljeve i uspoređivati različite količine podataka o modeliranju (Pomerantsev i Rodionova, 2018).
Ne postoji najbolja metoda klasifikacije. Svaki pojedini zadatak zahtjeva primjenu odgovarajuće kemometrijske metode, koja je najprikladnija za odgovor na postavljeno pitanje (Pomerantsev i Rodionova, 2018; Rodionova i sur., 2016).

2.5.3.1.1. Meko neovisno modeliranje analogne klase (SIMCA)

Meko neovisno modeliranje analogne klase (engl. *Soft Independent Modelling of Class Analogy*, SIMCA) je prva tehnika klasnog modeliranja uvedena u kemiji i danas je jedna od najpoznatijih klasifikacijskih metoda modeliranja. Jednoklasna je metoda koja se definira kao „meka“, jer nije postavljena hipoteza o distribuciji varijabli, odnosno klasifikacija je dvosmislena (meka) tj. svaki se uzorak može istovremeno dodijeliti u nekoliko klasa. Ova je metoda i neovisna, jer se svaka klasa modelira neovisno (Ballabio, 2009). SIMCA je ne-vjerojatnosna metoda modeliranja zasnovana na udaljenosti, a koju je uveo Svante Wold (1976). Kako je SIMCA jednoklasna metoda, objašnjavamo je koristeći se samo jednom klasom, koja je predstavljena matricom X dimenzija I (uzorci) s J (varijable). Cilj je razviti klasifikator (pravilo), koji odlučuje o novom uzorku x. Uzorak je ili prihvaćen (pripada klasi X) ili je odbijen (ne pripada klasi X). Da bi se taj problem riješio, matrica X se prikazuje kao oblak I točaka u J-dimenzionalnom prostoru varijabli (Slika 19). Ovaj oblak često ima specifičan oblik. Uzrokovane snažnom korelacijom između varijabli, točke se nalaze u blizini hiperravnine (potprostor) dimenzije $A < J$. Da bi se odredila dimenzija A i konstruirala ova hiperravnina, primjenjije se PCA (Pomerantsev, 2014). Dakle, SIMCA modeli se temelje na PC, koje su po definiciji smjerovi maksimalne varijance i, prema tome maksimalne informacije, u multivarijantnom podatkovnom prostoru (Jolliffe, 2002). PC se izračunavaju neovisno za svaku pojedinu klasu. PC definiraju takozvani SIMCA unutarnji prostor. Uzorci trening seta za modeliranje klase se projiciraju na PC unutarnjeg prostora, dobivajući vrijednosti faktorskih bodova za svaki uzorak na svakoj PC. Rasponi takvih PC faktorskih bodova definiraju model klase. Takav model ima oblik segmenta (jednodimenzionalni unutarnji prostor), pravokutnika (dvodimenzionalni unutarnji prostor), paralelepipeda ili hiperparalelepipeda (trodimenzionalni unutarnji prostor), obzirom na to da su PC ortogonalne po definiciji (Oliveri, 2017).



Slika 19. SIMCA klasifikacija (prilagođeno iz Wiberg, 2004).

Svaki element podataka se može predstaviti kao zbroj dvaju vektora. Jedan od vektora nalazi se na hiperravnini (projekcija), a drugi vektor je okomit na hiperravninu (rezidualni). Duljine ovih vektora važni su pokazatelji pripada li objekt klasi. Ove se duljine nazivaju bodovna udaljenost (engl. *score distance,* SD) *i* ortogonalna udaljenost (engl. *orthogonal distance*,OD). Čim se novi objekt smatra kandidatom koji pripada klasi, objekt se projicira na postojeći potprostor i izračunavaju se utjecajne vrijednosti i devijacija objekta. Uspoređujući ove vrijednosti s kritičnim (engl. *cutoff)* razinama, koje su utvrđene pomoću trening seta, moguće je odlučiti o pripadnosti novog objekta u klasu (Pomerantsev, 2014; Pomerantsev i Rodionova, 2020).

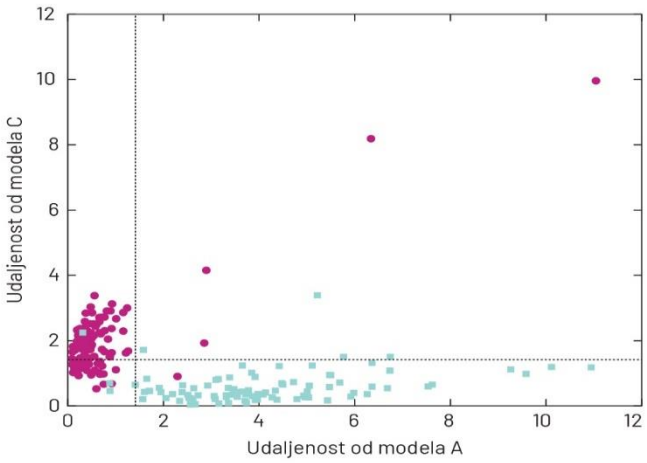
Kritična vrijednost ove udaljenosti, koja određuje prihvaćanje/odbijanje novog uzorka modelom, je definirana kritičnom vrijednošću Fisherove statistike na unaprijed određenoj razini pouzdanosti (engl. *confidence level*), s obzirom na to da predpostavlja da reziduali slijede multivarijantnu normalnu distribuciju (Slika 19.; Oliveri, 2017).

Praktična korisna primjena klasnog modela je strogo povezana sa njegovom pouzdanošću u predviđanju. Validacija modela, odnosno prediktivne sposobnosti novih uzoraka, koji se nisu koristili za formiranje modela, je ključna točka. Obično, validacijske strategije dijele uzorke na podskupove: trening set (ili kalibracijski set), koji se koristi za formiranje modela, i test set (ili evaluacijski set), koji se koristi za procjenu valjanosti modela. Setovi moraju sadržavati uzorke poznate klasne pripadnosti. Također, pouzdana validacija zahtjeva da se za izradu modela ne koriste informacije iz test seta uzoraka, kako bi se izbjegla precijenjena mogućnost predviđanja (Oliveri, 2017). Procjena prediktivne sposobnosti modela može se provesti bilo na jednom testnom skupu (postupak u jednom koraku) ili različitim evaluacijskim setovima, slijedeći iterativni postupak.

Kada se koristi jedan test set, odabere se obično manji set uzoraka koji čine test set, a preostali se koriste za formiranje trening seta. Raspodjela može biti proizvoljna, temeljena na slučajnom odabiru ili izvedena na način ujednačenog dizajna uzorkovanja, kao što su dizajn Kennard-Stone algoritmom i njegove modifikacije (Kennard i Stone, 1969). Ovi algoritmi generiraju dva podskupa uzoraka, koji istražuju cijelu domenu varijabilnosti, i ravnomjerno su raspoređeni unutar njega. Jedan od najčešćih izbora iterativnih validacijskih postupaka je postupak unakrsne validacije (engl. *cross.validation,* CV; Oliveri, 2017).

SIMCA rezultati mogu se grafički vizualizirati. Grafikoni PCA (grafikon opterećenja i grafikon faktorskih bodova), koji su izvedeni na trening setu, pružaju informacije o netipičnim uzorcima, podskupinama i unutarklasnoj strukturi. Koristan alat za interpretaciju SIMCA rezultata je Cooman dijagram, koji prikazuje diskriminaciju dviju klasa. Prikazom objekata ovim grafikonom lako je predočiti koliko je klasifikacija pouzdana (Slika 20; Berrueta i sur., 2007).

Nedostatak klasifikacijske SIMCA metode je formiranje modela s ciljem opisivanja varijacija unutar svake klase. Primjenom PCA na svaku klasu uzoraka pronalaze se smjerovi maksimalne varijance u prostoru klase, dok se ne pokušavaju pronaći smjerovi koji razdvajaju klase, kao npr. PLS-DA, koji izravno modelira klase na temelju deskriptora (Balabio i Todeschini, 2009)



Slika 20. Primjer Cooman dijagrama.

2.5.3.1.2. Regresija parcijalnih najmanjih kvadrata (PLS)

Regresija parcijalnih najmanjih kvadrata (engl. *Partial Last Squares*, PLS) najvažnija je regresijska metoda u kemometriji, koja se primjenjuje u raznim područjima kemije (analitička. fizikalna, kontrola industrijskih procesa i dr.) i koju je kasnih 60-ih godina prošlog stoljeća prvi

primjenio statističar Herman Wold (1975) u ekonometrijskim analizama, a njegov sin Svante (Wold i sur., 1984) prilagodio i primjenio za kemijsku primjenu nakon početne primjene Kowalskog i sur. (1986).

Modeliranje djelomičnih najmanjih kvadrata je multivarijatna projekcijska metoda za modeliranje odnosa između ovisnih varijabli (Y) i neovisnih varijabli (X). Princip PLS je pronaći komponente u ulaznoj matrici (X), koje opisuju što je više moguće relevantne informacije u ulaznim varijablama i u isto vrijeme imaju maksimalnu korelaciju s ciljnom vrijednosti Y, dajući manju težinu varijacijama koje su nebitne i izvor su šuma. Dakle, PLS modelira i X i Y istovremeno kako bi pronašao latentne varijable u X koje će predvidjeti latentne varijable u Y. PLS maksimizira kovarijancu između X i Y (Berrueta i sur., 2007).

PLS se može smatrati generalizacijom PCA. U PLS se istodobno provodi dekompozicija matrica X i Y (jednadžba 2.5.).

$$X = TP^t + E \qquad Y=UQ^t +F \tag{2.5}$$

gdje je:

X-ulazna matrica neovisnih vatijabli;

Y- matrica ovisnih varijabli;

T i U -matrice faktorskih bodova;

P i Q- matrice opterećenja;

E i F- matrice reziduala.

Projekcija se gradi kako bi se povećala korelacija između odgovarajućih vektora X-faktorskih bodova t_a i Y-faktorskih bodova u_a (Pomerantsev, 2014).

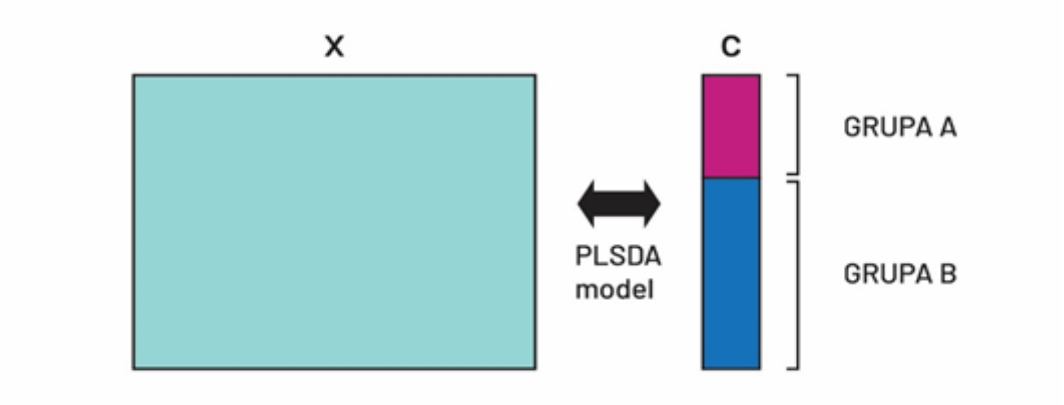
PLS faktorski bodovi interpretiraju se na sličan način kao i PCA faktorski bodovi. T-faktorski bodovi su nove koordinate točaka u X prostoru izračunate na takav način da zahvaćaju dio strukture u X, koji je za Y najpredvidljiviji. U faktorski bodovi sažimaju dio strukture u Y, koji je objašnjen s X uz danu komponentu modela. Odnos između T i U prikazuje odnos između X i Y za danu komponentu modela. P opterećenja izražavaju koliko svaka X varijabla doprinosi određenoj komponenti modela na sličan način kao i u PCA. Q opterećenja prikazuju direktan odnos između Y varijabli i T faktorskih bodova (Camo Analytics, 2014).

Važna značajka PLS je da uzima u obzir greške obje matrice, X i Y, i pretpostavlja da su jedanko distribuirane. PLS je pogodan za skupove s manje objekata od varijabli i visokim stupnjem međusobne korelacije između neovisnih varijabli. Postoji nekoliko algoritama za PLS, svaki s određenim prednostima, ovisno o istraživanom slučaju. Među njima nelinearni

algoritam djelomičnih najmanjih kvadrata (engl. *Nonlinear Iterative Partial Least Squares*, NIPALS) omogućuje izračunavanje glavnih komponenti jednu po jednu. Više informacija o algoritmima koji se koriste u PLS može se naći u dostupnoj literaturi (Berrueta i sur., 2007; Massart i Vandeginste, 1998; Rogers i Hopfinger, 1993; Arcos i sur., 1997).

2.5.3.1.3. Diskriminantna analiza parcijanih najmanjih kvadrata (PLS-DA)

Diskriminantna analiza parcijalnih najmanjih kvadrata (PLS-DA) je linearna klasifikacijska metoda, koja kombinira svojstva PLS regresije sa diskriminacijskom snagom klasifikacijskih tehnika i jedna je od diskriminacijskih tehnika koja se najviše primjenjuje. Metodu su uveli Barker i Rayens (2003). PLS-DA se temelji na PLS regresijskom algoritmu, koji traži latentne varijable s maksimalnom kovarijacijom s Y varijablama. PLS-DA primjenom regresije parcijalnih najmanjih kvadrata (PLS) pruža linearni graničnik (Wold i sur., 2001), koristeći binarne indekse pripadnosti klasi (npr. 0 i 1) za svaku klasu kao varijablu odgovora (Oliveri, 2017; Ballabio i Consonni, 2013). Model PLS-DA razvijen je pomoću PLS1 regresije kada se radi s jednom ovisnom varijablom i PLS2 u prisutnosti nekoliko ovisnih Y varijabli (Ballabio and Consonni, 2013), a koje su konstruirane između X, koja se koristi kao matrica prediktora i odgovora Y, koji je (I x K) lažna matrica (engl. *dummy matrix*). Svi elementi y_{ik} od Y su jedinice ili nule odabrani pravilom: Y_{ik} =1 ako uzorak i pripada klasi k; u suprotnom je y_{ik}=0.
Pravilo diskriminacije temelji se na usporedbi svakog retka predviđene matrice $\hat{Y}$ sa svakim vektorom odgovora uzorka. Uzorak i pripisuje se onoj klasi k koja je bliža. Da bi se procijenila udaljenost između uzorka i klase uzorka, matrica $\hat{Y}$ tretira se kao ulazni skup podataka za klasifikaciju. Međutim, to se ne može učiniti izravno, jer ova matrica ima rang K - 1, a odgovarajuća matrica kovarijance je singularna. Da bi se to pojedinačno riješilo, matrica $\hat{Y}$ se razlaže pomoću PCA, koja dimenziju smanjuje na K - 1. Matrica rezultata T predstavlja novi skup podataka, na koji se može primijeniti metoda klasifikacije (Pomerantsev i Rodionova, 2018). Matrice X i Y se rastavljaju u produkt druge dvije matrice bodova i opterećenja. Za razliku od PCA, koji koristi samo podatke matrice X, PLS također uzima u obzir podatke u matrici Y. Dakle, opterećenja X bloka izračunavaju se iz faktorskih bodova Y bloka, dok se opterećenja Y bloka određuju iz faktorskih bodova X bloka. Dekompozicije nisu neovisne, već su povezane kroz faktorske bodove blokova X i Y. Dakle, razvijen je model za svaku klasu. Što je element određenog stupca u Y bliži 1, a elementi ostalih stupaca 0, to je vjerojatnije da je objekt član određene klase (Slika 21; Berrueta i sur., 2007).



Slika 21. Prikaz PLS-DA modela za dvije klase. Vektor duljine C predstavlja numeričku oznaku za svaki uzorak prema članstvu u grupi (prilagođeno iz Brereton i Lloyd, 2014).

Glavna prednost PLS-DA je u tome što su relevantni izvori varijabilnosti podataka modelirani takozvanim latentnim varijablama (engl. *latent variables*, LV), koje su linearne kombinacije izvornih varijabli, i, slijedom toga, omogućuju grafičku vizualizaciju i razumijevanje različitih obrazaca među podacima te međusobnih odnosa među podacima pomoću LV faktorskih bodova i opterećenja. Opterećenja su koeficijenti varijabli u linearnim kombinacijama koje određuju LV, pa se stoga mogu protumačiti kao utjecaj svake varijable na svaki LV, dok faktorski bodovi predstavljaju koordinate uzoraka u hiperprostoru LV projekcije.
Kada su uključene više od dvije klase, PLS2 algoritam omogućuje predviđanje matrice varijabli odgovora, odnosno jednu matricu za svaku klasu. Kada broj varijabli znatno premaši broj objekata, PLS-DA u pravilu može pronaći graničnik (engl. *delimiter*), koji razlikuje dvije klase iako takve klase u stvarnosti nisu odvojene (Brereton i Lloyd, 2014). Stoga, u takvim slučajevima, osnova je temeljita validacija modela (Oliveri, 2017).

2.5.3.2. Formiranje, optimizacija i validacija klasifikacijskih modela

Cilj klasifikacijskih tehnika je pronaći matematički odnos između skupa opisnih varijabli (npr. kemijskih mjerenja) i kategoričkog odgovora. Identifikacija ovih odnosa između varijabli i vektora klase tada se može koristiti za predviđanje članstva nepoznatih uzoraka u jednoj od definiranih kategorija (klasa). Generalni kemometrijski pristup konstrukcije modela sastoji se od dvije važne faze. U prvoj fazi razvija se model koristeći trening set podataka. Idealno je da trening set uključuje sve moguće pa i buduće varijacije ciljne klase. U ovoj fazi konstruira se područje prihvaćanja za ciljnu klasu i uspostavljaju kriteriji prihvatljivosti. Druga faza je

validacija. Upotrebljava se za ispitivanje izvedbe razvijenog modela na uzorcima koji nisu korišteni kod formiranja modela. Za ispravnu provjeru valjanosti modela, test set bi trebao obuhvaćati uzorke ciljne klase i uzorke koji pripadaju alternativnim klasama (Rodionova i sur., 2014).

Obje metode, PLS-DA i SIMCA, temelje se na tehnikama projekcije, koje uključuju odabir složenosti modela, odnosno broj LV (PLS-DA) ili PC (SIMCA). Izbor ovog broja ima velik utjecaj na ishode modela. Ako je broj PC odnosno LV premalen, smanjuje se specifičnost modela, dok preveliki PC ili LV broj smanjuje osjetljivost modela (Rodionova i Pomerantsev, 2020). Kod provođenja optimizacije parametara modela, preporučljiva je strategija korištenja tri skupa uzoraka: trening seta, optimizacijskog seta i evaluacijskog set. Optimizacijski set koristi se za pronalaženje optimalnih postavki relevantnih parametara, dok se stvarna pouzdanost konačnog modela procjenjuje predviđanjem trećeg podskupa formiranog od objekata koji nisu utjecali ni na model ni na njegovu optimizaciju (Oliveri, 2017). Optimizacija čimbenika (predobrada podataka, parametri specifični za određenu metodu modeliranja) u potrazi za postavkom koja osigurava maksimalne performanse modela, dovodi i do značajnog rizika od precijenjene sposobnosti predviđanja modela. Precijenjena sposobnost predviđanja modela znači da model pretjerano odgovara zadanom podskupu uzoraka, koristeći značajan dio nebitnih podataka ugrađenih u analitičke podatke (npr. slučajni šum i neželjeni izvori varijacija). To obično dovodi do loše izvedbe u predviđanju modelom u stvarnim primjenama kod identifikacije nepoznatih uzoraka (Oliveri, 2017; Pomerantsev, 2014). Temeljna validacija modela je ključni korak za izbjegavanje precijenjene sposobnosti predviđanja modela i grešaka u predviđanjima na stvarnim uzorcima.

Postupkom validacije modela provjeravamo da li je formirani model dovoljno dobar da uspješno klasificira nepoznate uzorke, procjenom sposobnosti prepoznavanja i predviđanja modela. Idealna je situacija kada postoji dovoljno uzoraka za formiranje zasebnog trening seta, optimizacijskog seta i evaluacijskog seta, a svaki set sadrži reprezentativne uzorke za svaku klasu. Ovaj postupak validacije modela se naziva vanjska validacija. Kod ovog tipa validacije, test set je potpuno neovisan od postupka formiranja modela (odabir varijabli, procjena parametara, određivanje glavnih komponenti) (Berrueta i sur., 2007). U slučaju kada nije moguće formirati zaseban set uzoraka, provodi se postupak unakrsne validacije modela. Unakrsna validacija razdvaja N redova podatkovne matrice (uzoraka) u C skupine, prema unaprijed određenoj shemi (najčešće engl. *contiguous blocks* i *venetian blinds*). Model se izračunava C puta, i svaki put koristi jednu od grupa kao test set , a preostale uzorke kao trening set. C se obično kreće od 3 do N. N je ekstremni i obično pretjerano optimistični slučaj, koji je

općenito poznat kao „leave-one-out“ postupak (Oliveri, 2017). Postupkom unakrsne validacije sposobnost predviđanja modela određuje se razvojem modela s dijelom skupa podataka (trening set) i korištenjem drugog dijela podataka (test set) za testiranje modela. Oba seta - trening i test set, sadrže reprezentativne uzorke za svaku klasu. Ovaj postupak, koji se sastoji od razvoja modela i testiranja modela, ponavlja se nekoliko puta, tako da isti uzorci imaju vjerojatnost da se koriste i kao trening i kao test uzorci (Berrueta i sur., 2007).

Također je važan i odabir broja skupina za unakrsnu validaciju. Kad se skup podataka sastoji od malog broja uzoraka, preferira se veći broj grupa, kako bi se dobio veći dio uzoraka u trening setu i nekoliko uzoraka odabranih u svakoj validacijskoj grupi. Tako se na modelu formira mala preturbacija. Ako je broj uzoraka u skupu podataka relativno velik, preporučljivo je odabrati mali broj grupa za unakrsnu validaciju kako bi se odabralo više uzoraka za testiranje modela i izbjegla precjenjena sposobnost predviđanja modela (Ballabio i Consonni, 2013).

Procjena izvedbe klasifikacije temelji se na analizi takozvane matrice zbrke (engl. *confusion matrix*), koja prikazuje broj točnih i netočnih predviđanja za svaku klasu (Ballabio et al., 2018). Rezltati klasifikacije se tradicionalno opisuju izrazima „osjetljivost“, „specifičnost“ i „učinkovitost“ (Rodionova i Pomerantsev, 2020). Osjetljivost označava udio točno identificiranih uzoraka ciljne klase. Specifičnost je dio objekata alternativne klase koji su ispravno identificirani kao članovi te alternativne klase, odnosno udio ispravno odbijenih uzoraka koji nisu članovi ciljne klase (Oliveri, 2017). Učinkovitost je geometrijska sredina osjetljivosti i specifičnosti (Oliveri i Downey, 2012). Definicije osjetljivosti i specifičnosti se često baziraju na zapisima poput “istinski pozitivan“, „istinski negativan“ (Rodionova i Pomerantsev, 2020). Osjetljivost (engl. *sensitivity*) i specifičnost (engl. *specificity*) predstavljaju glavne validacijske parametre (engl. *figures of merit*) za procjenu kvalitete različitih identifikacijskih modela (Pomerantsev i Rodionova, 2018).

Klasna osjetljivost (engl. *class sensitivity;* CSNS) definirana je za svaku ciljnu klasu, kao postotak uzoraka klase koji su ispravno identificirani kao članovi te klase.

$$\mathrm{CSNS}(k) = n_{kk} / I_k \tag{2.6}$$

gdje je:

n_{kk} broj broj ispravno identificiranih *(*engl. *true positives*) članova klase za svaku klasu k, i

I_k je broj objekata/članova u klasi *k*.

Klasna specifičnost (engl. *class specificity;* CSPS) definirana je za svaku ciljnu klasu kao postotak uzoraka druge klase *(*engl. *non-target*), koji su ispravno prepoznati kao nekonzistentni sa ciljnim klasama.

$$\text{CSPS}\ (k) = 1 - \sum_{l \neq k} n_{kl} / \sum_{l \neq k} I_l \tag{2.7}$$

gdje je n_{kl} broj lažno pozitivnih za svaku klasu k i l.

Validacijski parametri (engl. *figures of merit,* FoM), koji karakteriziraju izvedbu cijelog identifikacijskog modela, su definirani kao:
ukupna osjetljivost (engl. *total sensitivity;* TSNS), koja se definira jednadžbom:

$$\text{TSNS} = \frac{1}{I} \sum_{k=1}^{K} n_{kk} \tag{2.8}$$

ukupna specifičnost (engl. *total specificity*; TSPS), koja je definirana donjom jednadžbom:

$$\text{TSPS} = 1 - \frac{1}{I} \sum_{k \neq l}^{K} n_{kl} \tag{2.9}$$

klasna učinkovitost (engl. *class efficiency;* CEFF(*k*)), koja karakterizira ukupnu kvalitetu identifikacije s obzirom na klasu k, i definirana je jednadžbom:

$$\text{CEFF(k)} = \sqrt{\text{CSNS}(k) \cdot \text{CSPS}(k)} \tag{2.10}$$

ukupna učinkovitost (engl. *total efficiency;* TEFF) karakterizira ukupnu kvalitetu identifikacije s obzirom na sve klase tj., i definirana je jednadžbom:

$$\text{TEFF} = \sqrt{\text{TSNS} \cdot \text{TSPS}} \tag{2.11}$$

Svi validacijski parametri (engl. *figures of merit*, FoM) izračunavaju se u odnosu na određenu klasu k, koja je uključena u klasifikaciju (Rodionova i Pomerantsev, 2020).

U PLS-DA metodi mogu se izračunati ukupne vrijednosti, koje karakteriziraju izvedbu cijelog modela. Ukupna osjetljivost (TSNS) definirana je kao stupanj svih istinski pozitivnih uzoraka. Ukupna specifičnost (TSPS) definirana je kao postotak svih istinski negativnih uzoraka. Ukupna učinkovitost (TEFF) geometrijska je sredina TSNS i TSPS.

U slučaju SIMCA (i ostalih metoda klasnog modeliranja) osjetljivost se dobiva za pojedinu ciljnu klasu pa je CSNS = TSNS. Specifičnost se može izračunati samo u prisutnosti alternativne klase. Ako postoji nekoliko neciljnih klasa, dobiva se specifičnost za svaku klasu posebno, kao CSPS (k). U skladu s tim, CEFF (k) i TEFF mogu se izračunati samo ako su predstavljene neciljne klase.

Prilikom klasifikacije uzoraka dvije su tipične greške koje se mogu pojaviti (Pomerantsev i Rodionova, 2020). Pri tom je potrebno odgovoriti na glavno pitanje, a to je što je "pozitivno" ?. Naravno, odgovor odmah definira i "negativno" kao nadopunu "pozitivnom". U statistici ovo znači izbor između hipoteze (pozitivne) i njezine alternative (negativne). Čim se odabere hipoteza vrijednost α (mjera pogrešnog odbijanja nulte, odnosno istinite hipoteze) ili omjer: α = broj lažno negativnih/broj svih istinitih uzoraka, naziva se pogreškom tipa I, a β vrijednost (mjera lažnog prihvaćanja alternativne hipoteze tj. hipoteze kada nije istinita) ili omjer: β = broj lažno pozitivnih / broj svih istinitih uzoraka, naziva se pogreškom tipa II. Ako se hipoteza i alternativa zamjene, pogreška tipa I postaje pogreška tipa II i obrnuto. Tijekom testiranja hipoteza (klasifikacije) važno je odlučiti koju vrstu pogrešaka treba minimizirati (Pomerantsev, 2014). Obje su pogreške štetne i generalno, za određeni skup podataka napor da se smanji jedna vrsta greške rezultira povećanjem druge vrste pogrešaka (Rodionova and Pomerantsev, 2020). Mogu li se istovremeno smanjiti obje greške? U principu je to moguće. Da bi se to učinilo, potrebno je promijeniti sam postupak donošenja odluka čineći ga učinkovitijim. Uobičajeni pristup bi bio povećanje broja varijabli koje kartakteriziraju objekte koji se klasificiraju. Stoga se metode klasifikacije, koje se koriste u kemometriji, uvijek temelje na multivarijantnim podacima (Pomerantsev, 2014).

U slučaju jednoklasne klasifikacije (engl. *one-class classification*) greška tipa I (α) se naziva nivo značajnosti (engl. *significance level*), a greška tipa II za taj tip klasifikacije je jednaka α-1 (Pomerantsev, 2014). U klasičnoj binarnoj diskriminaciji, kada su uključene dvije ekvivalentne klase, izbor koja je klasa pozitivna, a koja negativna potpuno je nevažan. To je samo slučajan izbor, koji ne utječe na ishod. Taj bi se slučaj mogao nazvati "simetrični" problem klasifikacije, jer bi se pozitivni i negativni događaji mogli zamjeniti (Rodionova i Pomerantsev, 2020).

3. MATERIJALI I METODE

3.1. Materijali

3.1.1. Uzorci

Prilikom izvedbe eksperimentalnog dijela ovog rada, korištene su komercijalne proizvodne serije pročišćenih meningokoknih polisaharida serogrupa (PMPS) A i C te eksperimentalne proizvodne serije PMPS W135 i Y. Sve serije proizvedene su u Imunološkom zavodu u Zagrebu.

3.1.2. Kemikalije i standardi

Korišteni su i visoko pročišćeni standardi PMPS A (NIBSC kod: 98/722) i PMPS C (NIBSC kod: 08/2014).

Prilikom izvedbe Ouchterloney metode korištene su slijedeće kemikalije i reagensi:

- Na-barbital ($C_8H_{11}N_2NaO_3$), Sigma-Aldrich, St Louis, MO, USA
- Barbital ($C_8H_{12}N_2O_3$), Sigma-Aldrich, St Louis, MO, USA
- NaN_3, Sigma-Aldrich, St Louis, MO, USA
- Agar, Kemika, Hrvatska
- NaCl, Kemika, Hrvatska
- Coomassie Brilliant Blue R-250, Sigma-Aldrich, St Louis, MO, USA
- Etanol, 96%, Kemika, Hrvatska
- Octena kiselina, Kemika, Hrvatska
- Antiserum protiv meningokoknog polisaharida A (koza), Imunološki zavod, Zagreb
- Antiserum protiv meningokoknog polisaharida C (magarac), Imunološki zavod, Zagreb
- Antiserum protiv meningokoknog polisaharida W135 (magarac), Imunološki zavod, Zagreb
- Antiserum protiv meningokoknog polisaharida Y (magarac), Imunološki zavod, Zagreb

3.1.3. Oprema

U pripravi uzoraka i reagensa za izvođenje Ouchterlony metode korištena je semi-mikro analitička vaga AT261 Delta Range (Mettler Toledo, Švicarska).

Sušenje imunodifuzijskih agar ploča provedeno je u sušioniku UF450 (Memmert, Buechenbach, Njemačka)

Spektri uzoraka pročišćenih meningokoknih polisaharida snimljena su pomoću Bruker MPA Series FT-NIR spectrometrom (Bruker Optik GmbH,Ettlingen,Germany).

Ramanovi spektri pročišćenih meningokoknih polisaharida snimani su na Ramanovom modulu Fourier-Transform (FT) PerkinElmer GX spektrometra sa pripadajućim softverom Spectrum v3.02, tvrtke PerkinElmer, Inc., Waltham, Massachusetts, United States. Ovaj uređaj koristi laser čvrstog stanja – kristal itrij-aluminijevog granata dopiran atomima trostruko ioniziranog neodimija (Nd:YAG) koji daje blisko infracrvenu pobudu valne duljine 1064 nm (9395 cm^{-1}).

3.2. Metode

3.2.1. Ouchterlony metoda

Kao referentna metoda za identifikaciju pročišćenih meningokoknih polisaharida korištena je dvostruka imunodifuzija Ouchterloney (Ouchterloney, 1962), koja se temelji na reakciji antigena i specifičnog antitijela u sloju agara, pri čemu nastaje precipitacijska vrpca.

Imunodifuzijske agar ploče pripravljene su prelijevanjem 1% w/v agaroznog gela (agaroza u barbitalnom puferu, pH 8.6) na staklene ploče i stavljanjem u vlažnu komoru pri 4 °C. Središnji zdenac i periferni zdenci izbušeni su bušačem prema shemi.

Optimalne koncentracije meningokoknih polisaharida (10 mg/ml) određene su eksperimentalno gdje je 40 µL antiseruma (interni magareći anti-meningokokni serogrupe C i kozji anti - meningokokni serogrupe A, Imunološki zavod d.d., Zagreb) stavljeno u središnji zdenac,a 10 µL referentnog antigena i otopine testnih uzoraka u periferne zdence. Ploče su inkubirane u vlažnoj komori na sobnoj temperaturi 48 sati. Provjerene su imunoprecipitacijske linije, formirane kada je koncentracija antigena dostigla ekvivalentnost s antitijelom.

Iznimno ako precipitacijske vrpce nisu dovoljno vidljive oboji se gel kao što je opisano dalje.

Nakon završene difuzije namoči se gel vodom, potom prekrije vlažnim filter papirom i iznad njega stavi još pet suhih filter papira. Na vrh se stavi staklena ploča i uteg od 1 kg. Pritiskanje se ponovi još dva puta uz promjenu suhog papira u intervalima od 3 min. Gel se ispire s 1 mol

L^{-1} otopinom NaCl (dva puta po 15 min), te se ponovi postupak prešanja. Nakon prešanja gel se potpuno osuši toplom ili hladnom strujom zraka pomoću sušila za kosu.
Posušeni gel prenese se u otopini za bojanje i ostavi 5 min., ispere se 1-2 puta u vodi uz potresanje. Odbojavanje se dalje provodi 10 min. u otopini za odbojavanje. Pritom nije potrebno mijenjati otopinu za odbojavanje.

3.2.2. Metode vibracijske spektroskopije

3.2.2.1. NIR spektroskopija

Spektri uzoraka pročišćenih meningokoknih polisaharida snimljena su pomoću Bruker MPA Series FT-NIR spectrometrom (Bruker Optik GmbH,Ettlingen,Germany). NIR spektri uzoraka snimljeni su difuznom refleksijom zračenja, (spektralno područje 9,990-3,695 cm-1; razlučivanje 16 cm^{-1}; 64 snimaka; detektor sulfidnog olova (PbS). Spektri su snimani OPUS softverom verzija 4.2 (Bruker Optik GmbH).
Svaki uzorak snimljen je u prozirnoj bočici ravnog dna, začepljen i izmjeren, pri čemu se uzorak miješao snažnim mućkanjem između ponavljanja.

3.2.2.2. Ramanova spektroskopija

Ramanovi spektri pročišćenih meningokoknih polisaharida snimani su na Ramanovom modulu Fourier-Transform (FT) PerkinElmer GX spektrometra sa pripadajućim softverom Spectrum v3.02, tvrtke PerkinElmer, Inc. Waltham, Massachusetts, United States, Medinski fakultet Sveučilišta u Zagrebu. Ovaj uređaj koristi laser čvrstog stanja – kristal itrij-aluminijevog granata dopiran atomima trostruko ioniziranog neodimija (Nd:YAG) koji daje blisko infracrvenu pobudu valne duljine 1064 nm (9395 cm^{-1}).
Za svaki uzorak načinjeno je 50 snimaka u području od 100 do 3500 cm^{-1} s rezolucijom spektrometra od 4 cm^{-1}. Snaga lasera bila je 500 mW.

3.2.2.3. Obrada NIR i Raman spektralnih podataka

Kemometrijska obrada spektara različitim matematičkim metodama provedena je u programu Unscrambler X verzija 10.4 (Camo Software AS, Oslo, Norveška). Korištene su različite tehnike predobrade spektara, kao primjerice prva i druga derivacija, korekcija višestrukog raspršenja metodom (MSC), standardna normalna varijata (SNV), Sawitzki-Golay glačanje, jedinična vektorska normalizacija, uklanjanje trenda, te su rezultati istih međusobno

uspoređivani. Odabrano je područje valnih brojeva NIR spektara je od ν = 7768 - 3695 cm^{-1}. a Raman spektara području od 100 do 3500 cm^{-1}.

3.2.2.4. Multivarijatna analiza spektralnih podataka

Podaci dobiveni vibracijskom spektroskopijom obrađeni su nenadgledanim i nadgledanim multivarijatnim metodama analize. Korištene nenadgledane metode u svrhu eksploratorne analize su PCA i hijerarhijska aglomerativna klaster analiza (Ward metoda), a nadgledane su se koristile SIMCA i PLS-DA. Svi obrađeni spektri analizirani multivarijatnim metodama analizirani su u programu Unscrambler X verzija 10.4 (Camo Software AS, Oslo, Norveška).

3.2.2.4.1. Analiza NIR spektralnih podataka

NIR spektri obrađeni su u programu Unscrambler X verzija 10.4 (Camo Software AS, Oslo, Norveška) u području valnih brojeva od ν = 7768-3695 cm^{-1}. Multivarijatne metode analize (PCA, SIMCA, PLS-DA) provedene su na spektrima prethodno obrađenim matematičkim metodama Savitzky-Golay glačanje 3.9 s drugom derivacijom, te uspoređene sa SIMCA i PLS-DA modelima provedenim na spektrima prethodno obrađenim matematičkim metodama Savitzky-Golay glačanje 3.9 s drugom derivacijom i SNV-om. Metoda klasteriranja s Ward algoritmom provedenim na spektrima prethodno obrađenim matematičkim metodama Savitzky-Golay glačanje 3.9 s drugom derivacijom i SNV-om kako bi se utvrdilo razdvajanje među grupama polisaharida.

Obrađeni spektri podvrgnuti su SIMCA metodi klasnog modeliranja i PLS-DA metodi diskriminacijske klasifikacije. Validacija oba modela provedena je validacijom iz nezavisnog skupa uzoraka koji nisu sudjelovali niti u kalibraciji niti u optimizaciji modela.

3.2.2.4.2. Analiza Raman spektralnih podataka

Ramanovi spektri obrađeni su u programu Unscrambler X verzija 10.4 (Camo Software AS, Oslo, Norveška) u području valnih brojeva od 3500 cm^{-1} do 100 cm^{-1}. Multivarijatne metode analize (PCA, SIMCA, PLS-DA) provedene su na prethodno obrađenim spektrima SNV i uklanjanjem trenda polinomom četvrtog stupnja. Kako bi se potvrdilo razdvajanje među grupama polisaharida, provedena je i hijerarhijska klasterska analiza Ward algoritmom Raman spektralnih podataka matematički obrađenih (SNV i uklanjanje trenda polinomom četvrtog stupnja) u spektralnom području $\tilde{\nu}$ = 3500 - 100 cm^{-1}.

Obrađeni spektri podvrgnuti su SIMCA metodi klasnog modeliranja i PLS-DA metodi diskriminacijske klasifikacije. . Optimizacija oba modela provedena je unakrsnom validacijom, a

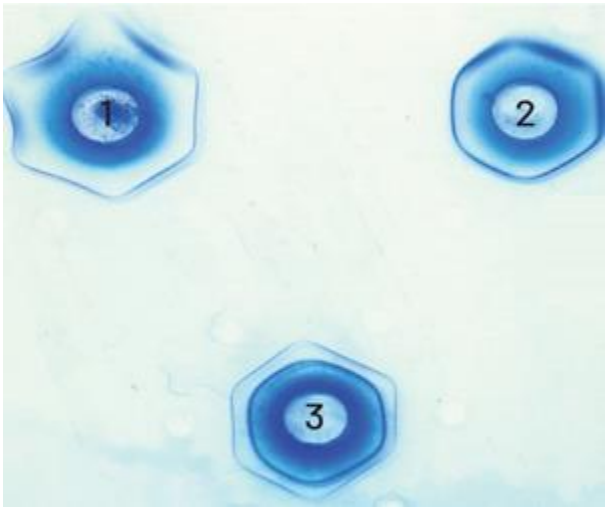
validacija modela, vanjskim postupkom korištenjem neovisnog seta uzoraka koji nisu sudjelovali u kalibraciji odnosno optimizaciji modela.

4. REZULTATI I RASPRAVA

Proizvodne serije uzoraka pročišćenih meningokoknih polisaharida serogrupa A i C, te eksperimentalne serije uzoraka meningokoknih polisaharida serogrupa W135 i Y, koje proizvodi Imunološki zavod, Zagreb, prikupile su se u razdoblju od dvije godine. Navedeni uzorci koristili su se za razvoj i validaciju novih metoda vibracijske spektroskopije i to NIR modela za identifikaciju PMPS A i PMPS C, te Raman modela za identifikaciju PMPS A i PMPS C. Pri razvoju NIR, odnosno Raman modela snimljeni spektri (NIR i Raman) uzoraka kemometrijski su se obradili različitim matematičkim metodama predobrade. U svrhu procjene strukture podataka napravljena je eksploracijska analiza glavnih komponenti te hijerarhijsko aglomerativno klasteriranje. NIR i Raman spektralni podaci raspodijeljeni su u setove te su formirani, optimirani i validirani SIMCA i PLS-DA modeli za identifikaciju uzoraka meningokoknih polisaharida serogrupe A i C. Prediktivna sposobnost modela procjenjena je validacijskim parametrima osjetljivost, specifičnost i učinkovitost. Kao referentna metoda korištena je metoda dvostruke imunodifuzije (Ouchterloney metoda).

4.1. Potvrda identiteta PMPS A, C, W135 i Y dvostrukom imunodifuzijom - Ouchterlony metodom

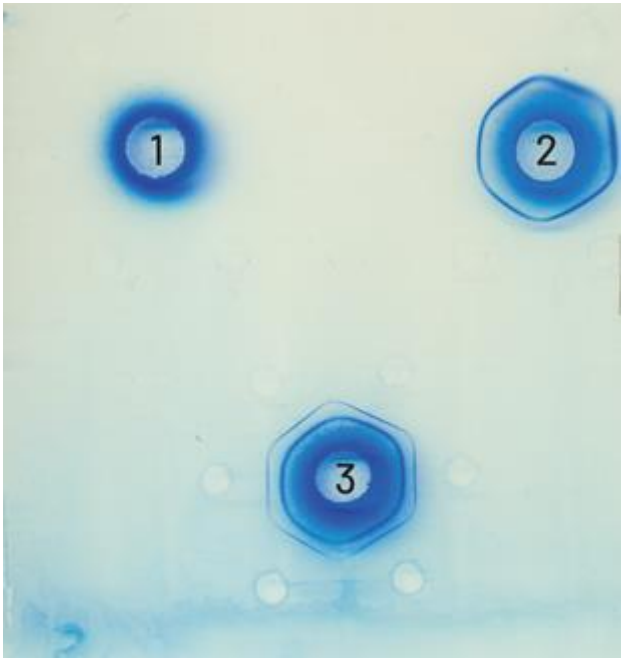
Prije početka izvođenja dvostruke imunodifuzijske tehnike (Ouchterlony metoda) , bilo ju je nužno optimirati, odnosno odrediti prikladne uvjete za precipitaciju eksperimentalnih serija PMPS W135 i PMPS Y. Da bi se odredio optimalni odnos koncentracija antigena i antitijela nužnih za precipitaciju, referentni antiserum za PMPS W135 i PMPS Y se nanio u centralni zdenac, a serija razrijeđenja antigena nanijela se u zdence koji okružuju centralni zdenac. Nanijela su se razrijeđenja 1:1, 1:2, 1:4, 1:8, 1:16 i 1:32. U prikazanom primjeru (Slika 22.) optimalna koncentracija antigena, s odgovarajućom koncentracijom antitijela, bila bi razrijeđenje 1:8 ili 1:16 ili 1: 32 na slici su navedena razrijeđenja najslabije naglašena.



Slika 22. Optimizacija Ouchterlony metode za određivanje identiteta PMPS W135 i PMPS Y. U zdencu 1 se nalazi referentni antiserum i serije razrijeđenja antigena PMPS W135 (1:1, 1:2, 1:4,1:8,1:16 i 1:32); u zdencu 2 nalazi se referentni antiserum i serije razrijeđenja antigena PMPS Y (1:1, 1:2, 1:4, 1:8, 1:16 i 1:32); a u zdencu 3 su referentni antiserumi PMPS W135 i Y i-laboratorijski priređeno-eksperimentalno cjepivo serogrupe W135 i Y.

Svim uzorcima proizvodnih serija PMPS A, PMPS C, te eksperimentalnih serija PMPS W135 i PMPS Y potvrđen je identitet referentnom metodom dvostruke imunodifuzije (Ouchterlony metodom).
Primjer pripravljene imunodifuzijske agar ploče sa formiranim imunoprecipitacijskim linijama koje potvrđuju identitet PMPS A i PMPS C prikazane su na Slici 23.

U središnje zdence stavljeno je 40 µL magarećeg anti-meningokoknog antiseruma serogrupe C (zdenac 1), 40 µL kozjeg anti-meningokoknog antriseruma serogrupe A (zdenac 2), a po 10 µL referentnog antigena serogrupe A i otopine testnih uzoraka PMPS A u periferne zdence. U zdenac 3 stavljeno je po 20 µL kozjeg anti-meningokoknog antriseruma serogrupe A i magarećeg anti-meningokoknog antiseruma serogrupe C, a u periferne zdence 10 µL cjepiva protiv meningokoka serogrupe A i C.



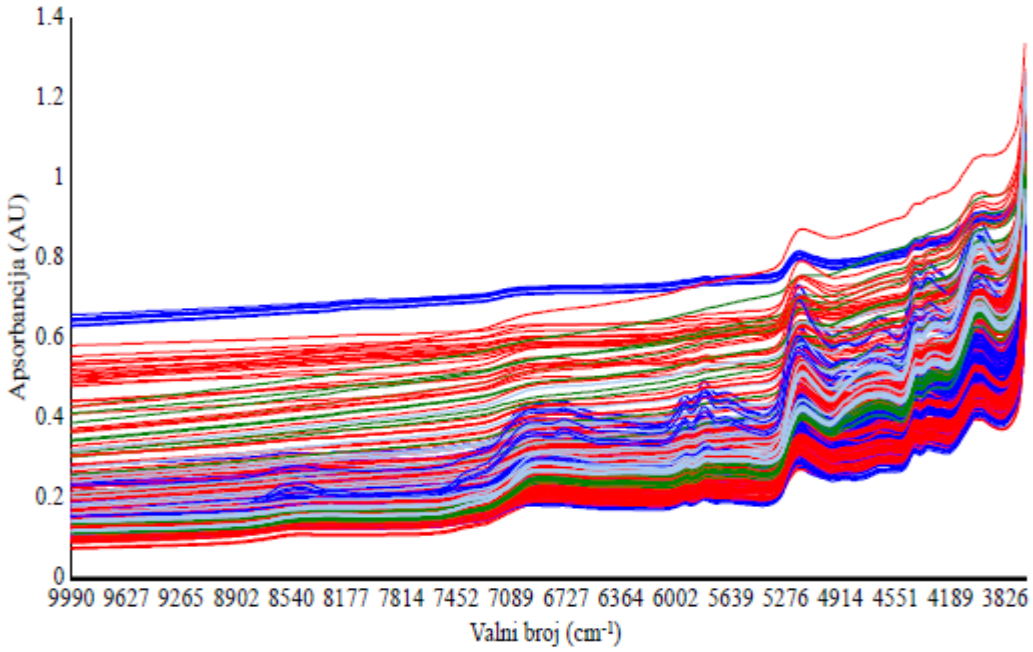
Slika 23. Određivanje identiteta PMPS A i PMPS C Ouchterlony metodom. U centralnom zdencu 1 nalazi se antiserum PMPS C, a u perifernim zdencima nalaze se referentni antigen PMPS A i testni uzorak PMPS A; u centralnom zdencu 2 nalazi se antiserum PMPS A a u perifernim zdencima nalaze se referentni antigen PMPS A i testni uzorak PMPS A, u centralnom zdencu 3 nalaze se antiserumi PMPS A i PMPS C, a u perifernim zdencima se nalazi uzorak cjepiva protiv meningokoka serogrupe A i C.

Na slici 23. jasno se vide formirane imunoprecipitacijske linije koje potvrđuju identitet ispitanih uzoraka PMPS A i PMPS C.

4.2. Razvoj i validacija NIR modela za identifikaciju PMPS A i PMPS C

4.2.1. NIR spektri PMPS A, C, W135 i Y

Ukupno su snimljena 589 NIR spektra PMPS A, C, W135 i Y. Pri tome je korišteno 30 serija PMPS A, 33 serije PMPS C te po jedna neovisna serija PMPS W135 i Y. Snimljeni spektri prikazani su na Slici 24.



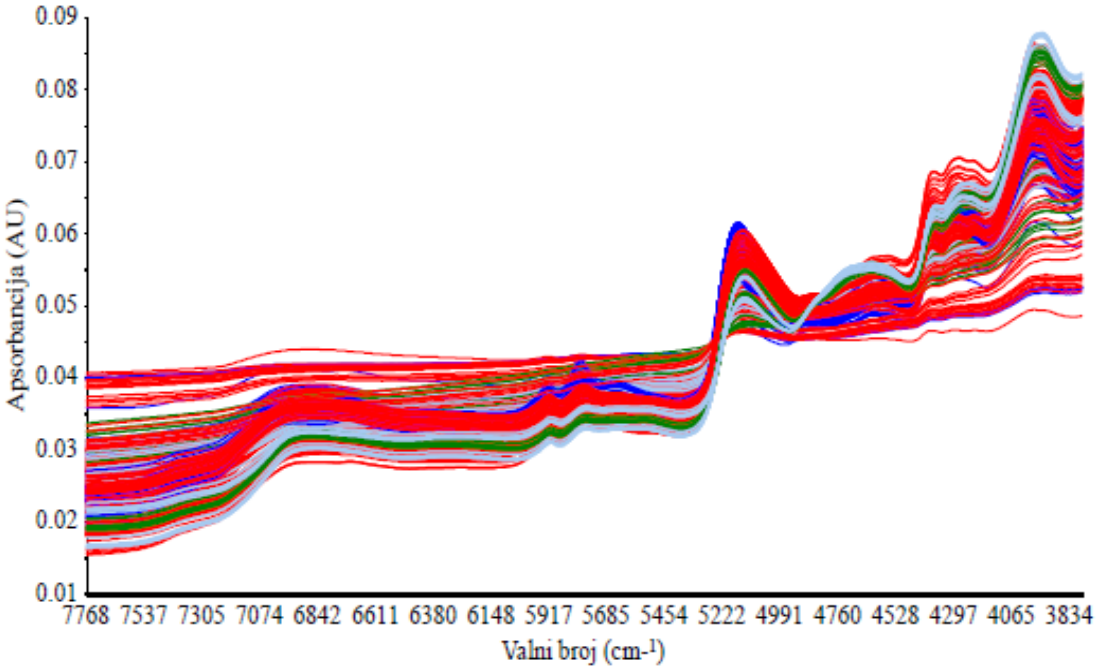
Slika 24. NIR spektri PMPS A (plavo), PMPS C (crveno), PMPS W135 (zeleno) i PMPS Y (sivo) u spektralnom području $\tilde{\nu}$ = 9990 - 3695 cm^{-1}.

4.2.2. Kemometrijska obrada snimljenih NIR spektara PMPS A, C, W135 i Y

Za kreiranje robustnog klasifikacijskog modela važno je osigurati spektralnu reproducibilnost. Varijacija između uzoraka pojedine serogrupe PMPS mora biti manja od varijacije među uzorcima različitih serogrupa PMPS. NIR spektralne podatke bilo je potrebno obraditi kako bi se maksimalno umanjile varijacije, koje nisu bitne za identifikaciju PMPS A i PMPS C. Cilj predobrade NIR spektralnih podataka je ukloniti podatke iz spektra koji se odnose na nedostatke nastale uslijed rasipanja svjetla, razne šumove, i fizikalna svojstva PMPS A i PMPS C koji nisu bitni za identifikaciju ovih polisaharida. Obradom NIR spektralnih podataka se poboljšava daljnja eksploracijska analiza ovih podataka kao i rezultirajući klasifikacijski NIR modeli za identifikaciju ovih polisaharida.

Kako bi se umanjili neželjeni učinci pomaka bazne linije NIR spektara i maksimizirali fini spektralni detalji, neobrađeni NIR spektri uzoraka PMPS A, C W135 i Y (Slika. 24.) su prethodno obrađeni različitim postupcima matematičke obrade prije daljnje analize ovih spektara, te su rezultati međusobno uspoređeni. Matematički obrađeni spektri PMPS A, PMPS C, PMPS W135 i PMPS Y prikazani su na Slikama 25. - 28.

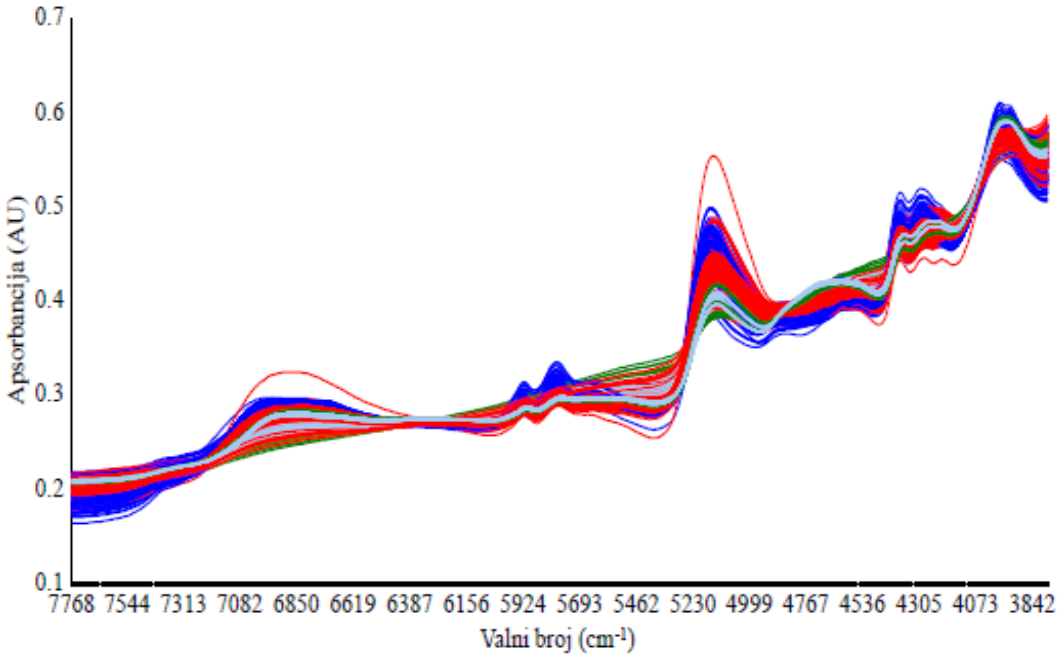
Kako bi se svi podaci dobili na približno istoj skali, NIR spektralni podaci su se transformirali jediničnom vektorskom normalizacijom (Slika 25.).



Slika 25. NIR spektri PMPS A (plavo), PMPS C (crveno), PMPS W135 (zeleno) i PMPS Y (sivo) matematički obrađeni jediničnom vektorskom normalizacijom u spektralnom području $\tilde{\nu}$ = 7768 - 3695 cm^{-1}.

Varijabilnost NIR spektara za sva četiri polisaharida (PMPS A, C, W135 i Y), koja se javlja uslijed rasipanja svjetla zbog razlike u veličini čestica uzoraka ispravljeno je višestrukom korekcijom raspršenog zračenja (MSC). Korekcija raspršenja je korištena za ispravljanje aditivnih i multiplikativnih učinaka karakterističnih fizikalno-kemijskih svojstava PMPS A, C, W135 i Y u NIR spektrima.

Na Slici 26. prikazani su NIR spektri za sva četiri polisaharida (PMPS A, C, W135 i Y), koji su matematički obrađeni MSC postupkom.

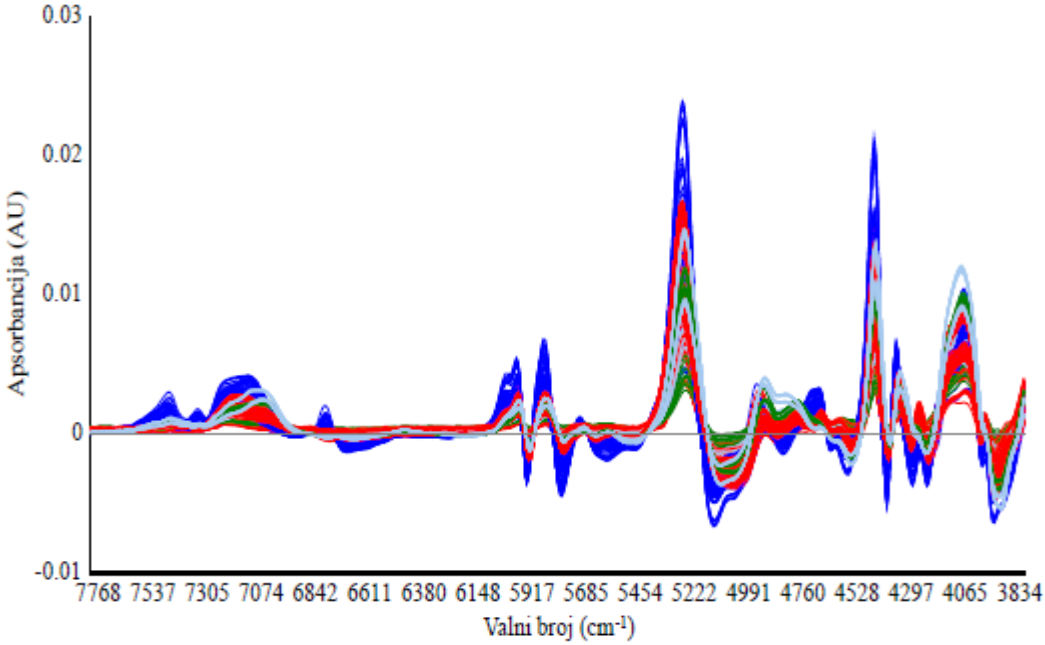


Slika 26. NIR spektri PMPS A (plavo), PMPS C (crveno), PMPS W135 (zeleno) i PMPS Y (sivo) matematički obrađeni višestrukom korekcijom raspršenog zračenja (MSC) u spektralnom području $\tilde{\nu}$ = 7768 - 3695 cm^{-1}.

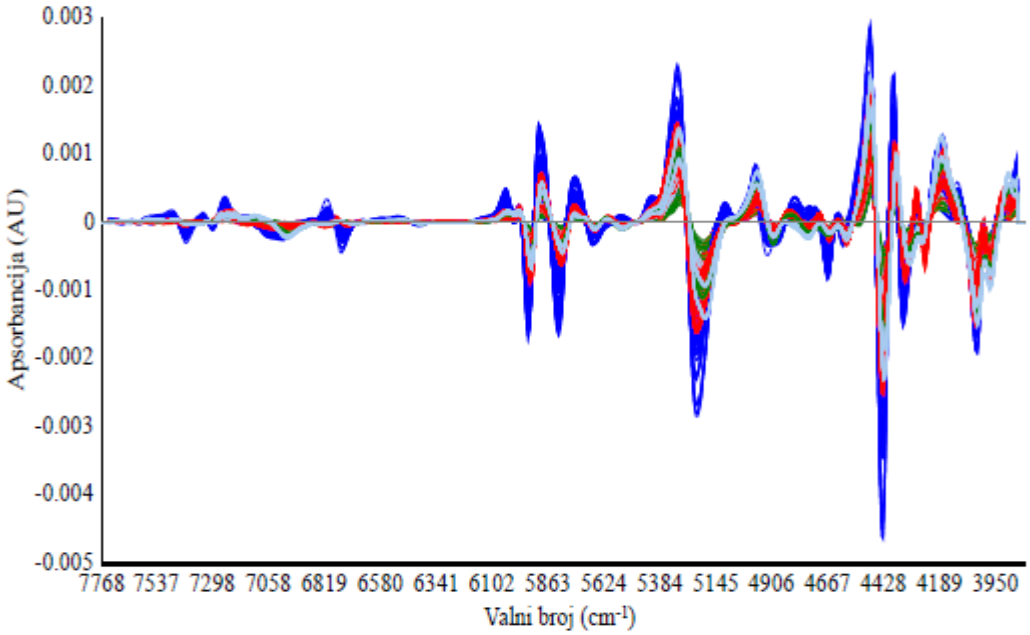
Predobrade NIR spektralnih podataka metodama derivacije pokazale su se efikasnima u korekciji aditivnih i multiplikativnih varijacija baznih linija NIR spektara (Poglavlje 2.7.1.). Pomoću prve i druge derivacije NIR spektralnih podataka može se ukloniti linearna pozadina spektara, dok signal (AU) pri bilo kojoj valnoj duljini tj. valnom broju ($\tilde{\nu}$) ostaje proporcionalan koncentraciji sastojaka polisaharida, baš kao što je bila i izvorna apsorpcijska vrpca NIR spektara.

Druga derivacija za obradu NIR spektara može se koristiti kao pomoć u rješavanju preklapajućih apsorpcijskih vrpci i uklanjanju pomaka bazne linije NIR spektara, a pri tome se ističu ključne spektralne razlike male apsorbancije pri određenim valnim duljinama.

Kako su derivacijama obrađeni NIR spektri osjetljivi na visokofrekventni šum, koji potiče od fizikalno-kemijskih karakteristika polisaharida, NIR spektri PMPS A, C, W135 i Y su također obrađeni primjenom Savitzky-Golay algoritma, kojim se polinomom trećeg stupnja interpoliraju vrijednosti absorbancije (AU) kroz devet eksperimentalnih spektralnih točaka. Dobiveni rezultati su prikazani na Slikama 27. i 28.

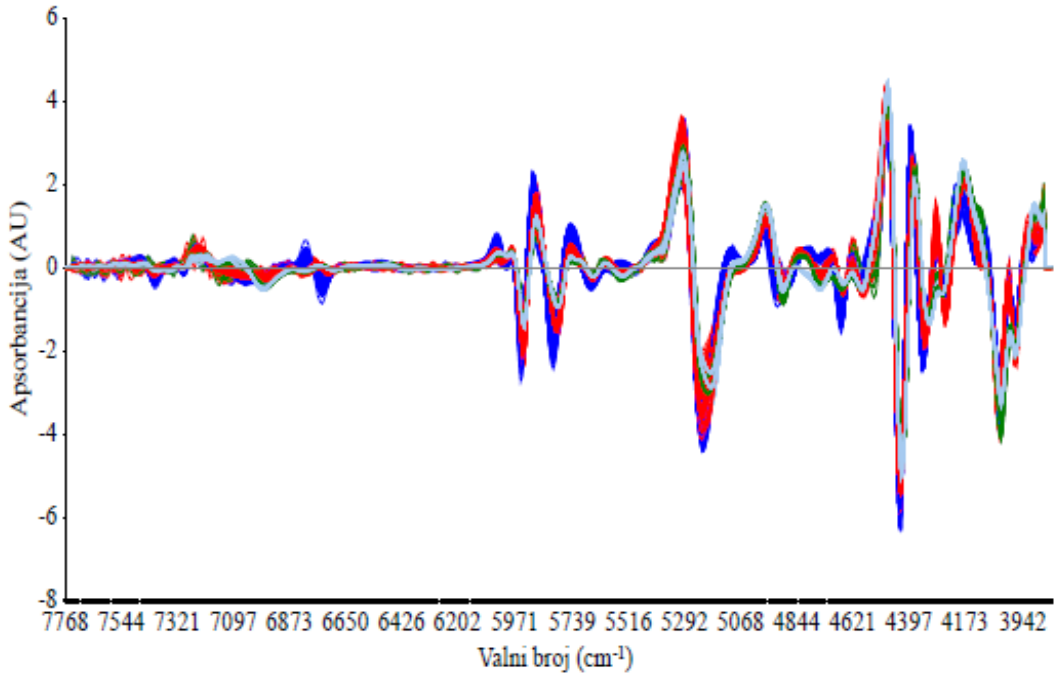


Slika 27. NIR spektri PMPS A (plavo), PMPS C (crveno), PMPS W135 (zeleno) i PMPS Y (sivo) matematički obrađeni Savitzky-Golay glačanjem 3.9 s prvom derivacijom u spektralnom području $\tilde{\nu}$ = 7768 - 3695 cm^{-1}.



Slika 28. NIR spektri PMPS A (plavo), PMPS C (crveno), PMPS W135 (zeleno) i PMPS Y (sivo) matematički obrađeni Savitzky-Golay glačanjem 3.9 s drugom derivacijom u spektralnom području $\tilde{\nu}$ = 7768 - 3695 cm^{-1}.

Dodatno, snimljeni NIR spektri PMPS A, C, W135 i Y su obrađeni standardnom normalnom varijatom (SNV) za skaliranje i centriranje NIR spektara, te kako bi se uklonila pozadina i raspršenje uzrokovano različitim veličinama čestica ovih polisaharida (Slika 29.).



Slika 29. NIR spektri PMPS A (plavo), PMPS C (crveno), PMPS W135 (zeleno) i PMPS Y (sivo) matematički obrađeni (SNV i Savitzky-Golay glačanje 3.9 s drugom derivacijom) u spektralnom području $\tilde{\nu}$ = 7768 - 3695 cm^{-1}.

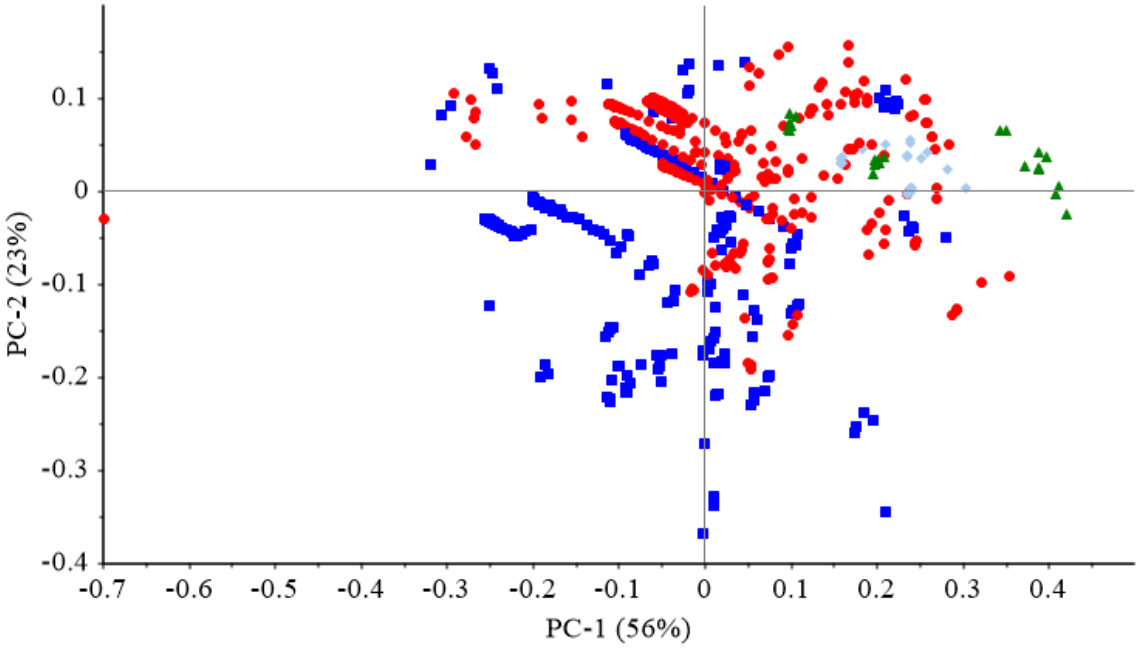
Kod ovako obrađenih (Savitzky-Golay glačanje 3.9 s drugom derivacijom i Savitzky-Golay glačanje 3.9 s drugom derivacijom i SNV-a) NIR spektara PMPS A, C, W135 i Y jasno se vidi međusobno razdvajanje ovih četiriju polisaharida, kao što je kasnije pokazano i kod analize glavnih komponenti (PCA; Slika 33. i 34.). Ovako obrađeni spektri koristit će se za daljnje formiranje modela.

Kako bi procijenili strukturu NIR spektralnih podataka, vizualizirali trendove među ovim podacima i detektirali grupe polisaharida PMPS A, C, W135 i Y, primijenjena je analiza glavnih komponenti (PCA).

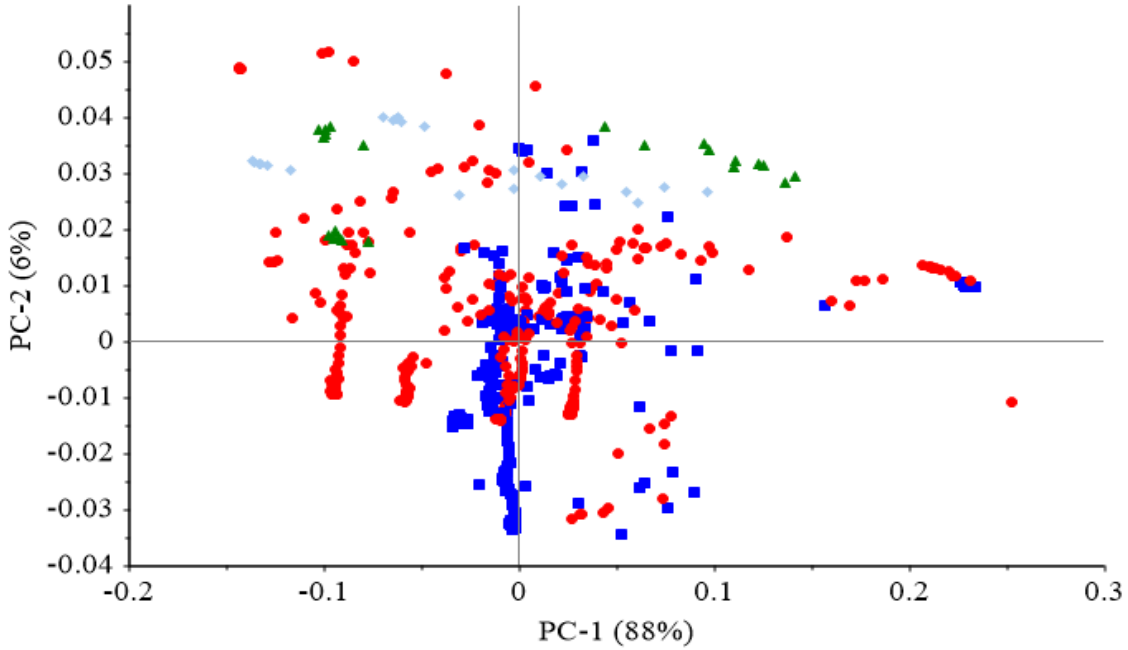
4.2.3. Eksploracijska analiza NIR spektralnih podataka PMPS A, C, W135 i Y

4.2.3.1. PCA

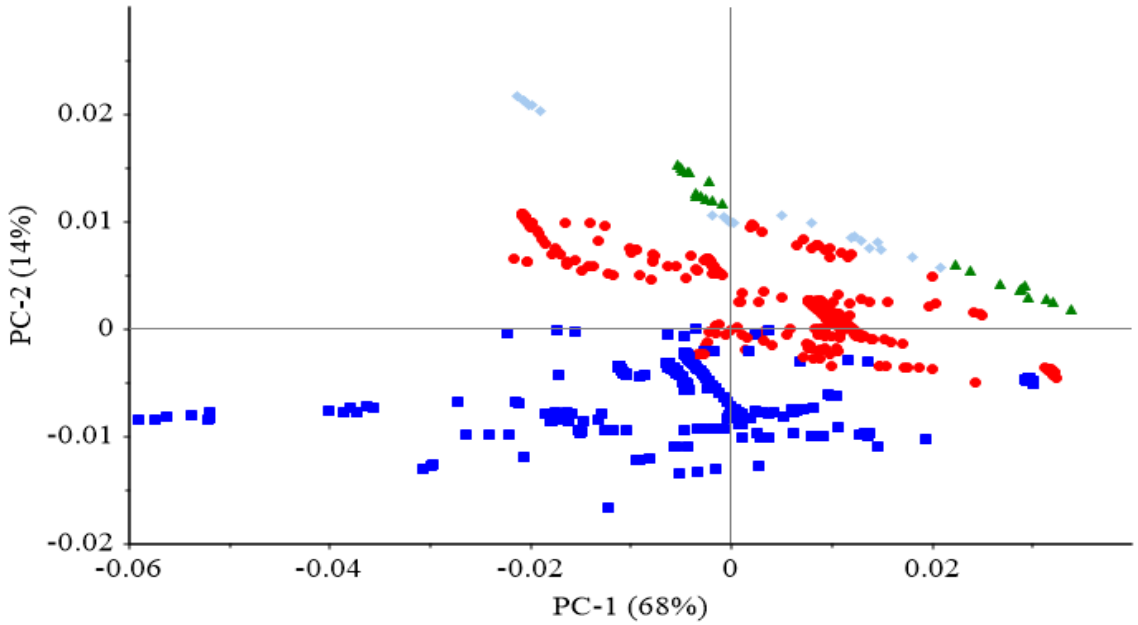
Primjenom PCA mogu se grupirati slični NIR spektralni podaci i to tako da se iz snimljenih NIR spektara ekstrahiraju relevantni podaci i istovremeno eliminiraju šumovi u ovim spektrima. PCA je napravljena kod NIR spektara PMPS A, C, W135 i Y, koji su prethodno obrađeni opisanim matematičkim metodama (Poglavlje 4.2.2.). Tako dobiveni podaci su vizualizirani pomoću glavnih komponenti, kako je to pokazano na Slikama 30. - 34.



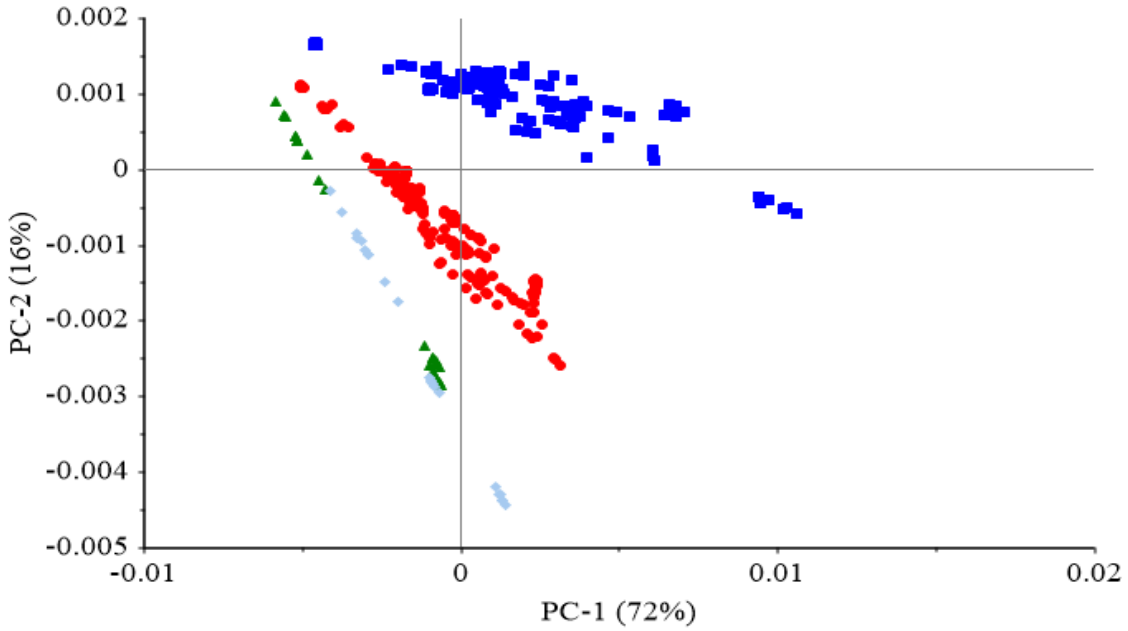
Slika 30. Raspodjela faktorskih bodova PC1 i PC2 matematički obrađenih (MSC) NIR spektara PMPS A, C, W135 i Y. Faktorski bodovi PC1 i PC2 se odnose na PMPS: A (plavi kvadrati), C (crveni krugovi), W135 (zeleni trokuti) i Y (sivi rombovi).



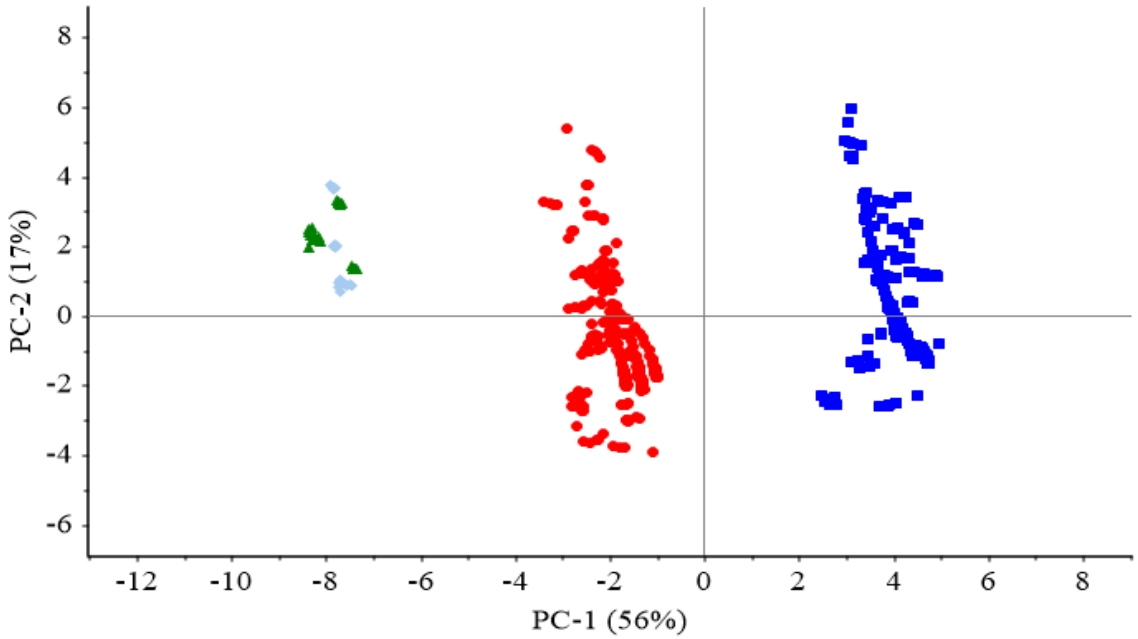
Slika 31. Raspodjela faktorskih bodova PC1 i PC2 matematički obrađenih (jediničnom vektorskom normalizacijom) NIR spektara PMPS A, C, W135 i Y. Faktorski bodovi PC1 i PC2 se odnose na PMPS: A (plavi kvadrati), C (crveni krugovi), W135 (zeleni trokuti) i Y (sivi rombovi).



Slika 32. Raspodjela faktorskih bodova PC1 i PC2 matematički obrađenih (Savitzky-Golay glačanje 3.9 s prvom derivacijom) NIR spektara PMPS A, C, W135 i Y. Faktorski bodovi PC1 i PC2 se odnose na PMPS: A (plavi kvadrati), C (crveni krugovi), W135 (zeleni trokuti) i Y (sivi rombovi).



Slika 33. Raspodjela faktorskih bodova PC1 i PC2 matematički obrađenih (Savitzky-Golay glačanje 3.9 s drugom derivacijom) NIR spektara PMPS A, C, W135 i Y. Faktorski bodovi PC1 i PC2 se odnose na PMPS A (plavi kvadrati), C (crveni krugovi), W135 (zeleni trokuti) i Y (sivi rombovi).

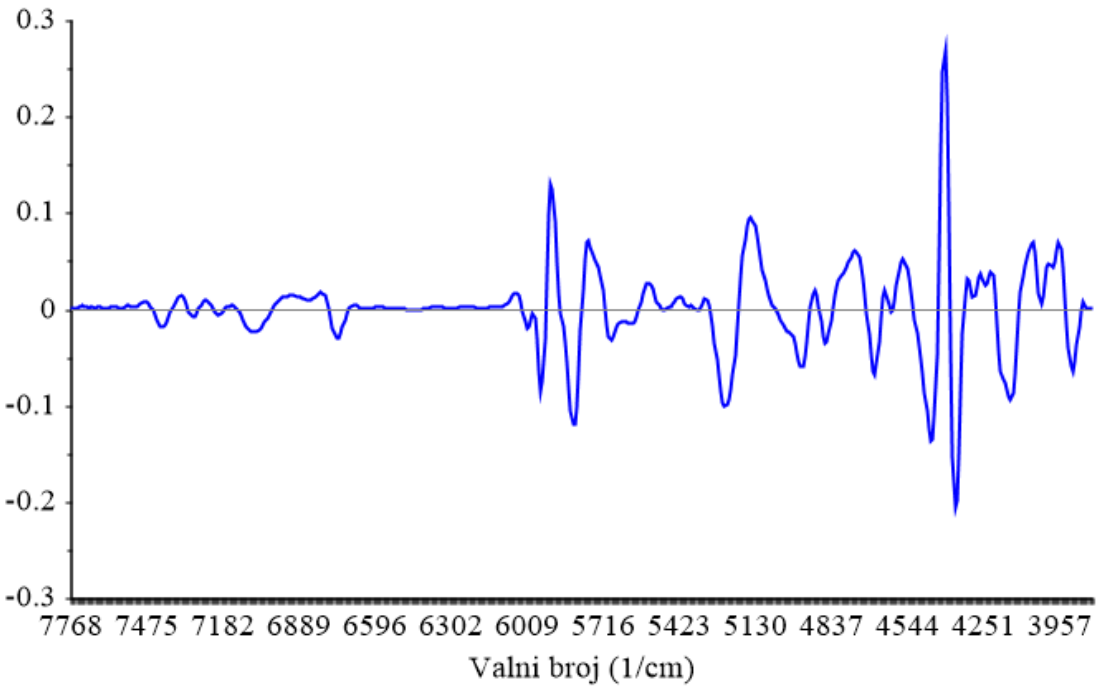


Slika 34. Raspodjela faktorskih bodova PC1 i PC2 matematički obrađenih (Savitzky-Golay glačanje 3.9 s drugom derivacijom i SNV) NIR spektara PMPS A, C, W135 i Y. Faktorski bodovi PC1 i PC2 se odnose na PMPS: A (plavi kvadrati), C (crveni krugovi), W135 (zeleni trokuti) i Y (sivi rombovi).

Dobiveni rezultati provedene eksploracijske analize matematički obrađenih NIR spektralnih podataka za PMPS A, C, W135 i Y procijenjeni su vizualnim pregledom odvojenosti ove četiri serogrupe pročišćenih meningokoknih polisaharida na odgovarajućim grafikonima faktorskih bodova (Slike 30. - 34.). Kod prve dvije glavne komponente - PC1 i PC2 (Slike 30. - 34.) jasno se vidi grupiranje i međusobne veze među uzorcima četiriju PMPS.

Nadalje je bilo potrebno načiniti profil opterećenja (poglavlje 2.5.2.1.2.), kako bi se definirale najvažnije NIR spektralne regije koje su ključne za međusobno razdvajanje ovih četiriju polisaharida. Osim toga, metodom hijerarhijskog klasteriranja (poglavlje 2.5.2.2.) grupirani su slični NIR spektri PMPS A, C, W135 i Y.

Profili opterećenja PMPS A, C, W135 i Y prikazani su na Slici 35.



Slika 35. PC 1 opterećenja po valnim brojevima ($\tilde{\nu}$) dobivena PCA analizom matematički obrađenih (Savitzky-Golay glačanje 3.9 s drugom derivacijom i SNV) NIR spektara PMPS A, C, W135 i Y.

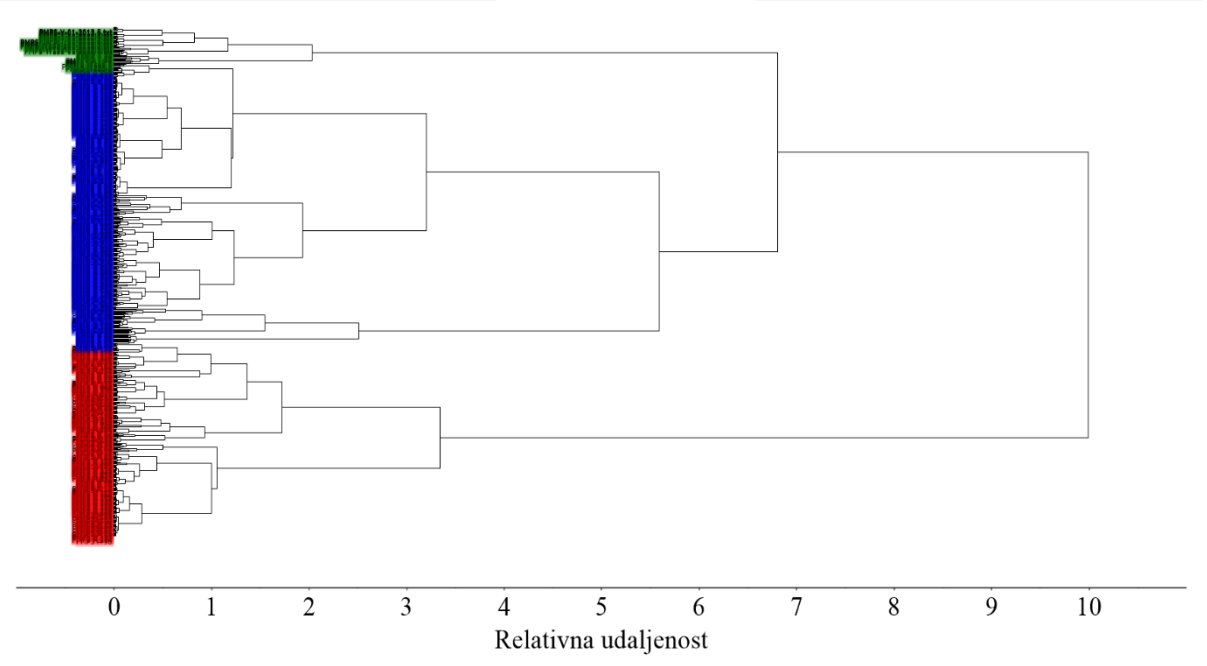
Izradom profila opterećenja za PC1 identificirane su najutjecajnije spektralne regije u NIR spektrima PMPS A, C, W135 i Y, na PC1 i odgovorne su za, diferenciranje četiri različite serogrupe meningoknih polisaharida.

Karakteristična spektralna regija u rasponu valnih brojeva $\tilde{\nu}$ = 5970 - 5910 cm^{-1} proizlazi iz prvog višeg tona acetamid metil (C-H) asimetričnog istezanja; zatim $\tilde{\nu}$ = 5780 - 5840 cm^{-1}

nastaje iz prvog višeg tona metilen (C-H) asimetričnog istezanja; spektralna regija $\tilde{\nu}$ = 5150-5240 cm^{-1} odgovara kombinaciji O-H istezanja, H-O-H deformacija i O-H savijanja; dok spektralna regija $\tilde{\nu}$ = 4360 - 4450 cm^{-1} odgovara kombinaciji C-H istezanja i CH_2 deformacije; a spektralna regija $\tilde{\nu}$ = 4060 - 4150 cm^{-1} odgovara kombinaciji istezanja C- H, C-C istezanja i C-O-C istezanja (Workman, 2001).

4.2.3.2. Klasterska analiza

PCA eksploratorna analiza podataka (poglavlje 4.2.3.1.) koristi se prvenstveno za određivanje opće povezanosti NIR spektralnih podataka. Međutim, dodatno treba utvrditi da li se uzorci PMPS A, C, W135 i Y mogu razvrstati u četiri različite skupine polisaharida. U svrhu provjere sličnosti uzoraka PMPS A, C, W135 i Y primijenjena je i metoda klasteriranja s Ward algoritmom. Rezultat cjelokupne klasterske analiza podataka s Ward algoritmom prikazan je kao dendogram za ove odabrane serije uzoraka PMPS A, C, W135 i Y (Slika 36.).



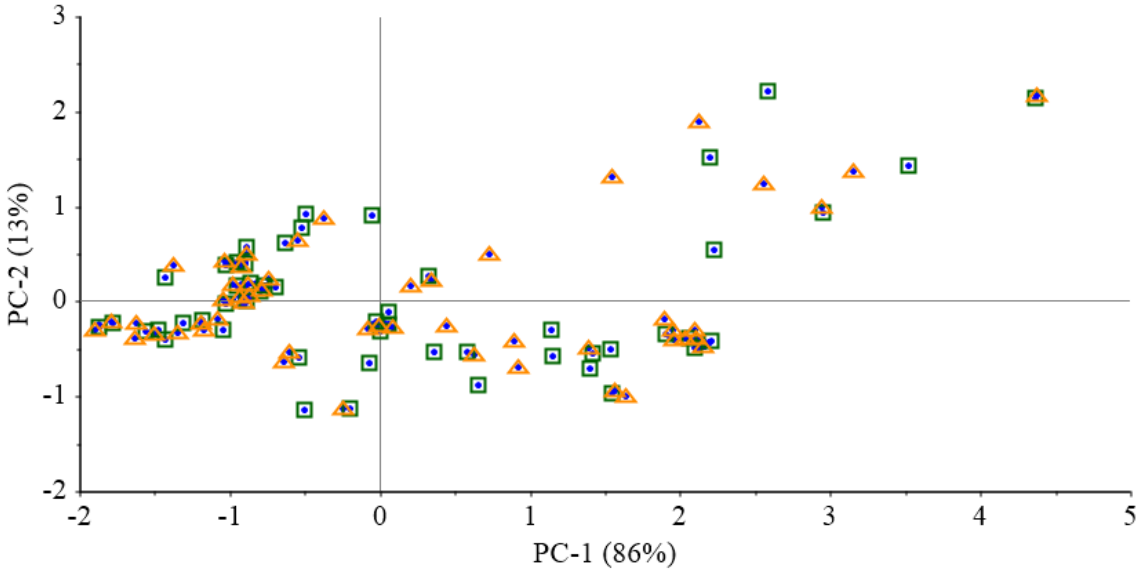
Slika 36. Dendografska klasifikacija uzoraka PMPS A, C, W135 i Y dobivena Ward algoritmom za hijerarhijsko grupiranje podataka u spektralnom području $\tilde{\nu}$ = 7768 - 3695^{-1}.

Iz Slike 36. mogu se identificirati tri tipa različitih uzoraka polisaharida. Klaster 1 (crveno) grupirao je uzorke PMPS A, klaster 2 (plavo) uzorke PMPS C dok je klaster 3 (zeleno) grupirao

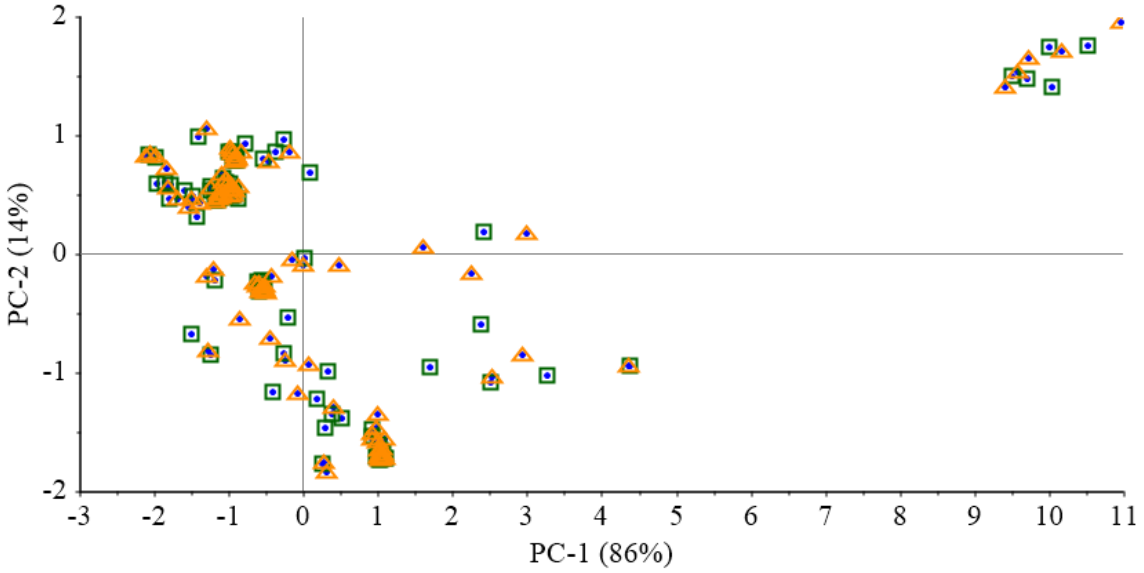
uzorke W135 i Y. Visina dendrograma na kojem su spojena dva klastera predstavlja udaljenost između dva klastera u podatkovnom prostoru, odnosno sličnost između dva klastera. Klasteri 2 (plavo) i 3 (zeleno) su povezani manjom udaljenošću veze, što ukazuje da imaju veću sličnost, što odgovara kemijskoj strukturi uzoraka PMPS C (plavo) i W135 i Y (oba zeleno). Klaster 1 (PMPS A, crveno) je manje sličan preostalim dvama klasterima, jer ima veliku udaljenost od klastera 2 (PMPS C, plavo) i 3 (W135 i Y, zeleno).

4.2.4. Raspodjela NIR spektara PMPS A, C, W135 i Y u kalibracijski, optimizacijski i evaluacijski set

Praktična primjena NIR modela, s pomoću kojega se planiralo utvrditi identitet PMPS A i C, usko je povezana sa pouzdanošću predviđanja identiteta nepoznatih uzoraka PMPS A i C. Validacija NIR modela, odnosno procjena sposobnosti predviđanja NIR modela da identificira nepoznate uzorke polisaharida A i C, je ključni zahtjev pri formiranju ovoga NIR modela.
Sukladno strategiji validacije NIR modela (poglavlje 2.5.3.2.), koja predviđa podjelu raspoloživih uzoraka na podskupove, uzorci polisaharida A i C raspodijeljeni su na tri seta: set za trening (ili kalibracijski set), koji se koristi za formiranje modela; test set 1 (ili optimizacijski set), koji se koristi za optimizaciju odnosno procjenu ispravnosti formiranog NIR modela; i treći set - evaluacijski set (ili vanjski validacijski set), koji koristi uzorke polisaharida A, C, W135 i Y, koji nisu sudjelovali u formiranju i optimizaciji NIR modela. Sva tri seta (set za trening ili kalibracijski set i test set 1 ili optimizacijski set, te evaluacijski ili vanjski validacijski set) moraju sadržavati uzorke poznate pripadnosti serogrupi. Validacija NIR modela zahtjeva da se pri formiranju NIR modela ne koriste uzorci iz test seta1, niti vanjskog validacijskog seta kako bi se izbjegla precijenjena sposobnost predviđanja ovog modela u budućoj uporabi.
Na Slikama 37 i 38. prikazana je podjela uzoraka PMPS A i PMPS C na trening (ili kalibracijski set) i test set 1 (ili optimizacijski set). Uzorci su podijeljeni na temelju analize njihove udaljenosti u trodimenzionalnom PCA prostoru prema uniformnom Kennard-Stone algoritmu (Slika 37 i Slika 38.). Kennard-Stone algoritam je prilagođen primjeni u analitičkoj kemiji, jer omogućuje da model formiran od trening seta (ili kalibracijskog seta) uzoraka pokriva većinu izvora varijacija unutar skupa podataka, i na taj način osigurava da model bude reprezentativniji za cijeli skup podataka (Kennard i Stone, 1969).



Slika 37. Podjela uzoraka PMPS A Kennard-Stone algoritmom na trening set i test set 1. Trening set (zeleni kvadrati), test set 1 (narančasti trokuti).



Slika 38. Podjela uzoraka PMPS C Kennard-Stone algoritmom na trening i test set1. Trening set (zeleni kvadrati), test set 1 (narančasti trokuti).

Na Slikama 37 i 38. mogu se vidjeti jednoliko raspodijeljeni reprezentativne skupovi podataka za trening i za test set polisaharida i to kroz cjelokupni skup podataka, kako bi se održala raznolikost kako u jednom tako i u drugom skupu uzoraka. Dodatno u test setu 1 PMPS A i PMPS C nalaze se i PMPS W135, PMPS Y te NIBSC standardi.

4.2.4.1. Raspodjela NIR spektara PMPS A u kalibracijski i optimizacijski set

Ukupno 172 NIR spektra od 20 serija PMPS A i 5 NIR spektara standarda za polisaharid A (NIBSC standard, kod: 98/722) Svjetske zdravstvene organizacije (WHO) podijeljeni su na (1) trening set (ili kalibracijski set), koji je sadržavao ukupno 111 snimljenih NIR spektara od 13 serija PMPS A, i (2) test set 1 (ili optimizacijski set), koji je sadržavao 61 NIR spektar snimljen kod sedam proizvodnih serija PMPSA i 5 NIR spektara za standard polisaharida A.

4.2.4.2. Raspodjela NIR spektara PMPS C u kalibracijski i optimizacijski set

Ukupno 251 NIR spektar od 23 serije PMPS C i 5 NIR spektara standarda za polisaharid C (NIBSC standard, kod: 08/2014) podijeljeni su na: (1) trening set, koji se sastojao od 153 snimljena NIR spektra od 16 serija PMPS C, i (2) test set 1, koji se sastoji od 98 NIR spektara od sedam serija PMPS C i 5 NIR spektara za standard polisaharida C.
Deset replikata od po jedne serije polisaharida W135 i Y su korišteni kao negativna proba za polisaharide A i C.
U svrhu evaluacije formiranog NIR modela korišten je treći set - evaluacijski set (poglavlje 4.2.4.3.).

4.2.4.3. Raspodjela NIR spektara PMPS A, C, W135 i Y u vanjski validacijski set

Ukupno 60 NIR spektara od 10 serija PMPS A, zatim ukupno 62 NIR spektra od 10 serija PMPS C te po 12 NIR spektara od po dvije serije PMPS W135 i Y (negativne probe) su korištene kao vanjski validacijski set budućeg NIR modela. NIR spektri vanjskog validacijskog seta nisu sudjelovali niti u formiranju niti u optimiranju PMPS A i C NIR modela.

4.2.5. NIR SIMCA model

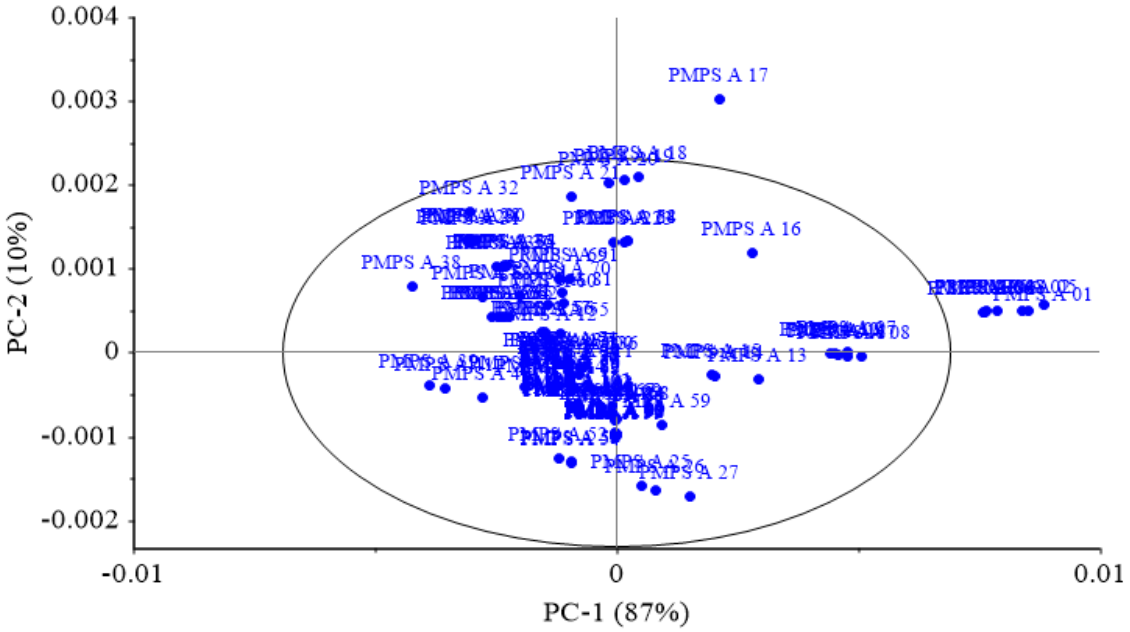
Kao što je već prije opisano u Općem dijelu (poglavlje 2.5.3.1.1.), NIR SIMCA model pripada grupi klasnog modeliranja i prikladan je za modeliranje kada je dostupan reprezentativan set uzoraka bez nekontroliranih varijacija za svaku pojedinu klasu (serogrupu) meningokoknih polisaharida. U SIMCA-i se PCA model fromira za svaku pojedinu klasu unutar skupa podataka. U ovom radu svaka pojedina serogrupa meningokoknih polisaharida modelirala se zasebno i to jedna po jedna. Svaka se klasa serogrupe modelirala na temelju sličnosti uzoraka PMPS unutar pojedine klase metodom definiranja prostora klase. Iznimno važno za SIMCA model je određivanje optimalne dimenzionalnosti modela, odnosno broja PC-a, što se utvrdilo formiranjem PCA svake pojedine klase (poglavlje 4.2.5.1.) te optimizacijom SIMCA modela.

4.2.5.1. PCA modeliranje NIR spektara PMPS A i C (Savitzky - Golay glačanje 3.9 s drugom derivacijom)

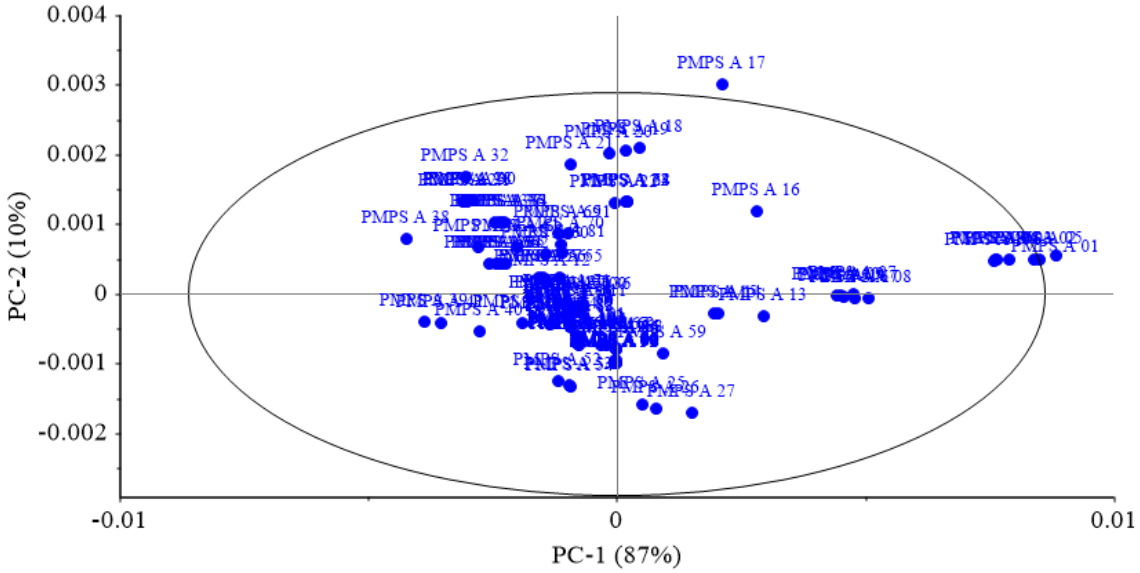
U NIR spektralnim podacima svaki valni broj ($\tilde{\nu}$) predstavlja jednu neovisnu dimenziju tj. varijablu, pa posljedično dimenzije skupa NIR spektralnih podataka postaju velike, a sposobnost razlučivosti grupe polisaharida/klasa - ograničena. Kako je opisano u Općem dijelu (poglavlje 2.5.2.1.), PCA se koristi u svrhu smanjenja dimenzija skupa NIR spektralnih podataka. Kod multidimenzionalnih podataka sličnosti i različitosti među podacima su teško uočljive, a to znači da ih je nemoguće grafički prikazati, pa nam PCA služi u svrhu smanjenja dimenzija i mogućnosti grafičkog prikazivanja multidimenzionalnih podataka.
Jedan od glavnih ciljeva PCA je identifikacije glavnih komponenti (PC), koje su od ključnog značaja za objašnjenje ukupne varijance NIR spektralnih podataka za svaki PMPS. Svaka je glavna komponenta PCA karakterizirana vrijednostima faktorskih bodova, opterećenja, reziduala uzoraka i uzoraka visokog utjecaja (utjecajnih vrijednosti). U ovom su radu provedene pojedinačne analize glavnih komponenti (PCA) NIR spektralnih podataka za svaki meningokokni polisaharid, posebno za PMPS A, a posebno za PMPS C.

4.2.5.1.1. Analiza glavnih komponenti NIR spektralnih podataka za PMPS A

PCA je provedena na kalibracijskom skupu uzoraka PMPS A, koji je obuhvatio 111 NIR spektara. NIR spektri, koji nisu bili dio kalibracijskog seta, nisu korišteni u ovom PCA, već su se koristili za optimizaciju i za validaciju formiranog NIR modela.



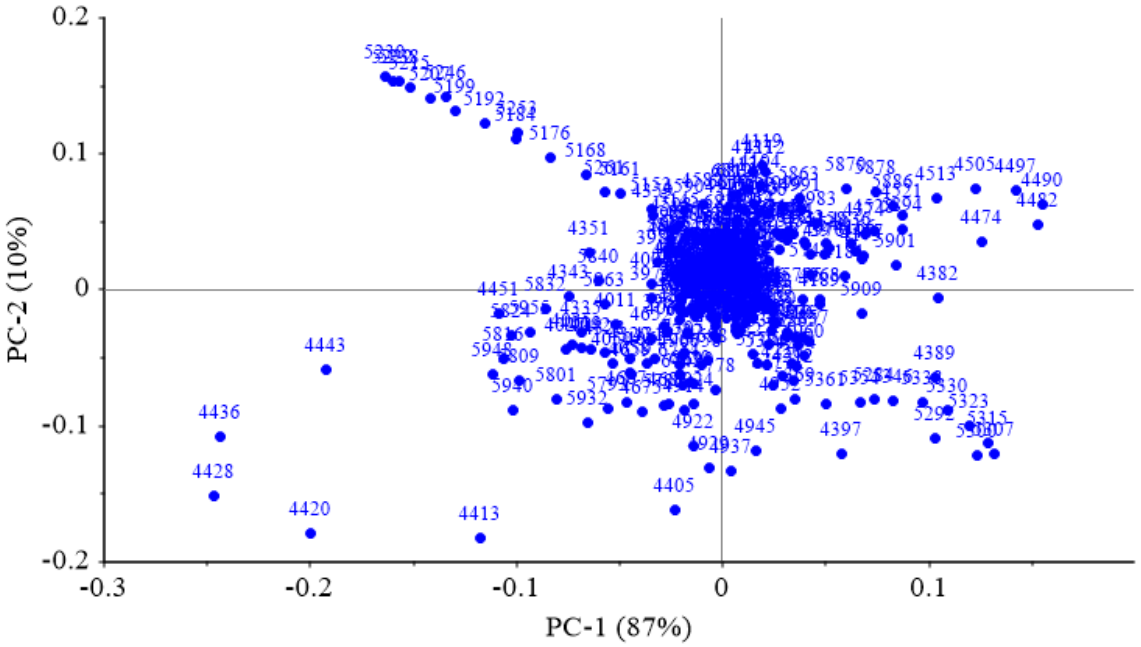
Slika 39. Raspodjela faktorskih bodova PC1 i PC2 matematički obrađenih (Savitzky-Golay glačanje 3.9 s drugom derivacijom) NIR spektara PMPS A u spektralnom području $\tilde{\nu}$ = 7768 - 3695 cm^{-1}. Elipsa označava 95 %-tni interval pouzanosti (Hotelling T^2 statistika).



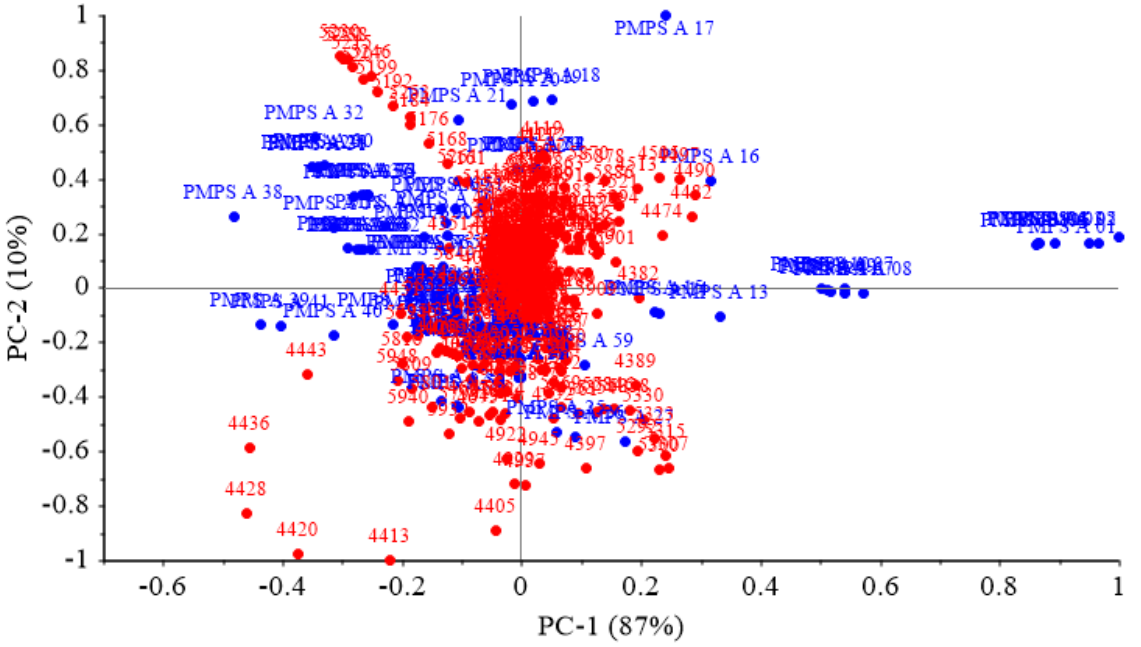
Slika 40. Raspodjela faktorskih bodova PC1 i P 2 matematički obrađenih (Savitzky-Golay glačanje 3.9 s drugom derivacijom) NIR spektara PMPS A u spektralnom području $\tilde{\nu}$ = 7768 - 3695 cm^{-1}. Elipsa označava 99 %-tni interval pouzanosti (Hotelling T^2 statistika).

Na Slikama 39. i 40. se može vidjeti jednolika raspodjela faktorskih bodova dobivenih za uzorke PMPS A kroz cijelo spektralno područje ($\tilde{\nu}$ = 7768 - 3695 cm^{-1}). Ovi prikazi faktorskih bodova pružaju uvid u međusobne odnose među uzorcima ovoga polisaharida (PMPS A). Prva glavna komponenta (PC1) oduhvaća 87% varijance, a druga glavna komponenta (PC2) 10% varijance. Uzorci polisaharida PMPS A 01-06, i PMPS A 17 nalaze se izvan Hotelling T^2 elipse (95% interval pouzdanosti) te ih treba, kao potencijalne netipične vrijednosti, dodatno istražiti statističkim metodama.

Također nas je zanimalo i (1) koje varijable su odgovorne za obrazac grupiranja među uzorcima, (2) koje varijable imaju najviše utjecaja na njihovo međusobno grupiranje odnosno razdvajanje i (3) kako varijable međusobno koreliraju. Slike 41. i 42. prikazuju odnose između svih varijabli NIR podataka za PMPS A istovremeno. Varijable koje pridonose sličnim informacijama su grupirane zajedno.



Slika 41. PC1 i PC2 opterećenje matematički obrađenih (Savitzky-Golay glačanje 3.9 s drugom derivacijom) NIR spektara PMPS A u spektralnom području $\tilde{\nu}$ = 7768 - 3695 cm^{-1}.



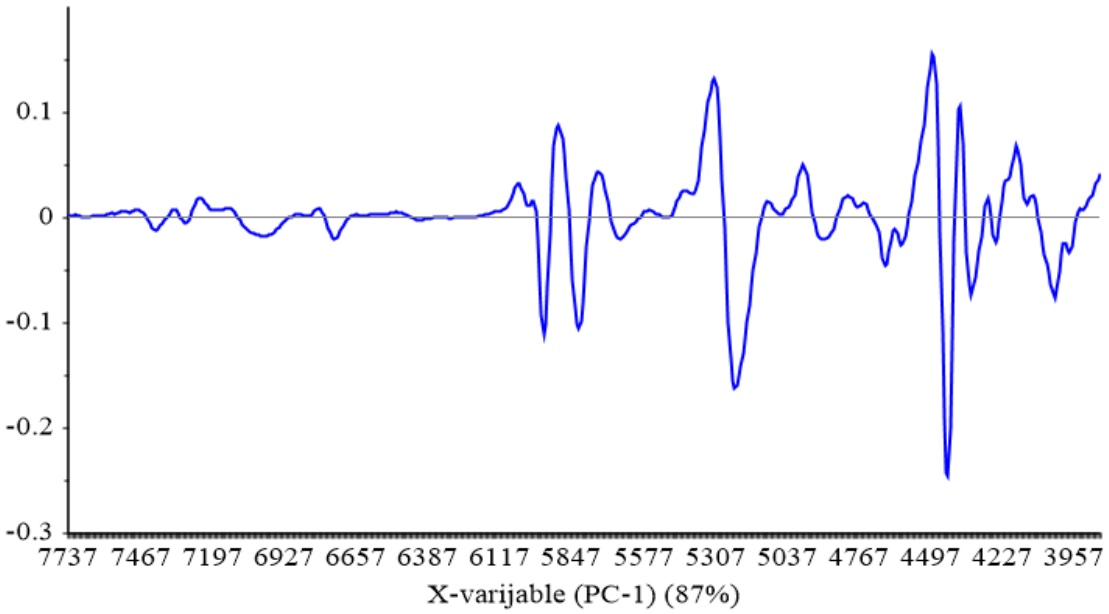
Slika 42. Raspodjela faktorskih bodova PC1 i PC2 matematički obrađenih (Savitzky-Golay glačanje 3.9 s drugom derivacijom) NIR spektara PMPS A u spektralnom području $\tilde{\nu}$ = 7768 - 3695 cm^{-1} zajedno sa PC1 i PC2 opterećenjem. Opterećenja - crveno, faktorski bodovi - plavo.

Svaka podatkovna točka faktorskih bodova (Slike 39. i 40.) predstavlja po jedan NIR spektar PMPS A. Svaka točka u odgovarajućoj slici opterećenja (Slike 41. - 42., označeno crvenom bojom) predstavlja po jedan valni broj unutar svakoga snimljenog NIR spektra. Ukoliko se želi identificirati vrpca koja je uzrok razlike između određenog NIR spektra od ostalih NIR spektara, bilo je potrebno usporediti opterećenja za svaki faktorski bod. Opterećenje za svaki uzorak tj. faktorski bod definira smjer razdvajanja tog određenog NIR spektra od ostalih NIR spektara za PMPS A.

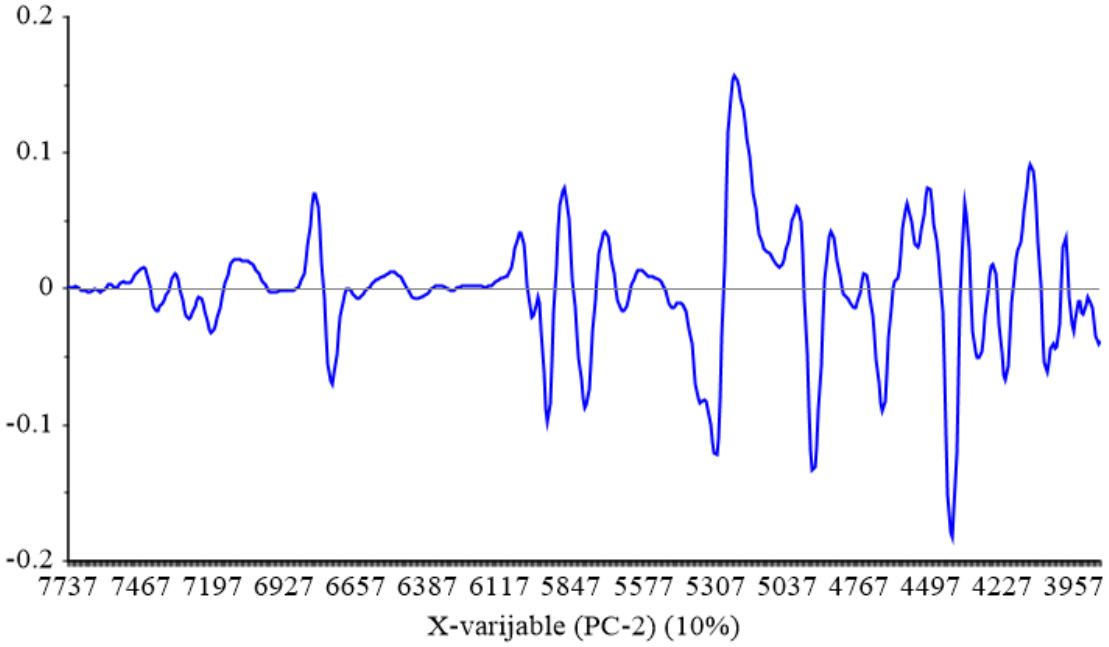
Slike s opterećenjima (Slike 41. - 42., označeno crvenom bojom) također pokazuju koji valni brojevi ($\tilde{\nu}$) koreliraju jedan s drugim i to kroz određivanje kuteva, koji nastaju kada se formira linija od jednog valnog broja do izvorišta i drugog relevantnog valnog broja do izvorišta. Što je ovako definiran kut manji između ovih formiranih linija, veća je korelacija između dvaju odabranih valnih brojeva. Mali kut podrazumijeva pozitivnu korelaciju između dvaju odabranih valnih brojeva, dok veći kut sugerira negativnu korelaciju između dvaju odabranih valnih brojeva. Kut od 90° ukazuje da nema korelacije između dvaju odabranih valnih brojeva tj. dvije karakteristike unutar NIR spektra.

Slika s raspodjelom faktorskih bodova zajedno s faktorskim opterećenjem (Slika 42.) je vrlo koristan prikaz za karakteriziranje opažanja korelacije među dobivenim NIR spektralnim podacima. Iz ove slike (Slika 42.) možemo utvrditi položaje faktorskih bodova i varijabli te njihov utjecaj na PC.

Na Slikama 43. i 44. se može vidjeti koji valni broj ($\tilde{\nu}$) koliko doprinosi pojedinoj glavnoj komponenti (PC1 i PC2).



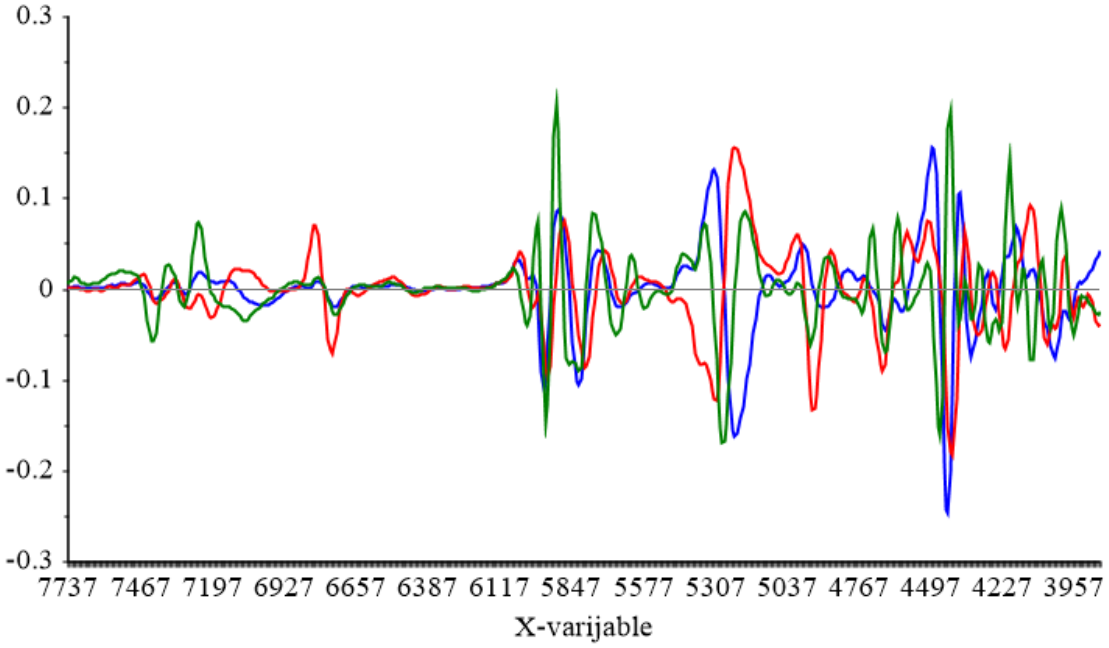
Slika 43. PC1 opterećenja po valnim brojevima ($\tilde{\nu}$) dobivena PCA analizom NIR spektara PMPS A.



Slika 44. PC2 opterećenja po valnim brojevima ($\tilde{\nu}$) dobivena PCA analizom NIR spektara PMPS A.

Valni brojevi ($\tilde{\nu}$) s najvećim opterećenjima najviše doprinose pojedinoj komponenti (PC1 i PC2). Preklopljena su opterećenja za PC1, PC2 i PC3 kako bi jasno prikazali i utvrdili koje su

NIR spektralne regije odgovorne za definiranje pojedine glavne komponente, odnosno koje su varijable koje imaju najveći utjecaj na model (Slika 45.).

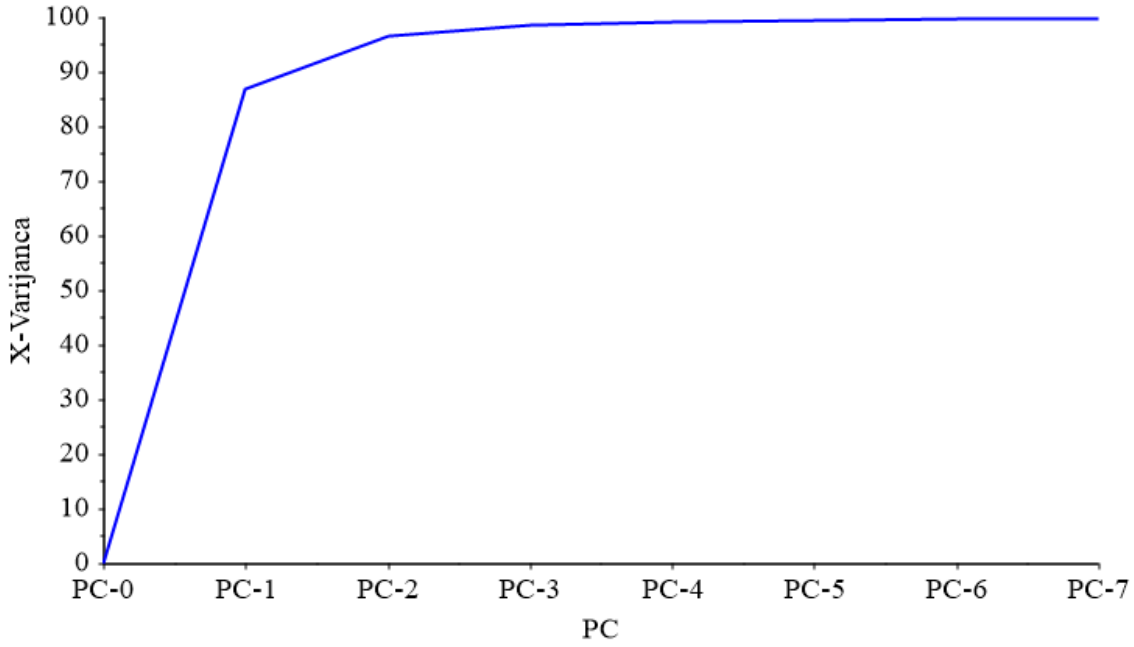


Slika 45. PC1 (plava), PC2 (crvena) i PC3 (zelena) opterećenja po valnim brojevima ($\tilde{\nu}$). Opterećenja su dobivena PCA analizom za skup od 111 NIR spektara PMPS A.

Ovim načinom utvrdio se doprinos svake varijeble na PC model, odnosno, identificirane su varijable koje imaju najveći utjecaj na pojedinu glavnu komponentu. Opterećenja prikazana na Slikama 43.-45. ukazuju na najvažnije spektralne regije, koje definiraju PC1 i PC2. Spektralna područja odgovorna za definiranje prve glavne komponente (PC 1; Slika 43.) su: $\tilde{\nu}$ = 5970.-.5910 cm^{-1} koje proizlazi od prvog višeg tona acetamid metil C–H asimetričnog istezanja; $\tilde{\nu}$ = 5780.–.5840 cm^{-1} proizlazi od prvog višeg tona metilen C–H asimetričnog istezanja; spektralno područje $\tilde{\nu}$= 5300.-.5100 cm^{-1} odgovara O-H kombinacijskim vibracijama; spektralno područje $\tilde{\nu}$ = 5150–5240 cm^{-1} odgovara kombinaciji polisaharidnog O-H istezanja, H-O-H deformacije i O-H savijanja; spektralno područje $\tilde{\nu}$ = 4900 - 4500 cm^{-1} proizlazi iz drugog višeg tona C=O istezanja, C-N istezanja i N-H savijanja u ravnini; spektralno područje $\tilde{\nu}$ = 4360 - 4450 cm^{-1} odgovara kombinaciji C–H istezanja i CH_2 deformacije. Spektralna područja odgovorna za definiranje druge glavne komponente (PC 2; Slika 44.) su: $\tilde{\nu}$ = 7000 - 6500 cm^{-1} proizlazi od prvog višeg tona O-H istezanja $\tilde{\nu}$ = 5970 - 5910 cm^{-1} proizlazi od prvog višeg tona acetamid metil C–H asimetričnog istezanja; $\tilde{\nu}$ = 5780 - 5840 cm^{-1} proizlazi od prvog višeg tona metilen

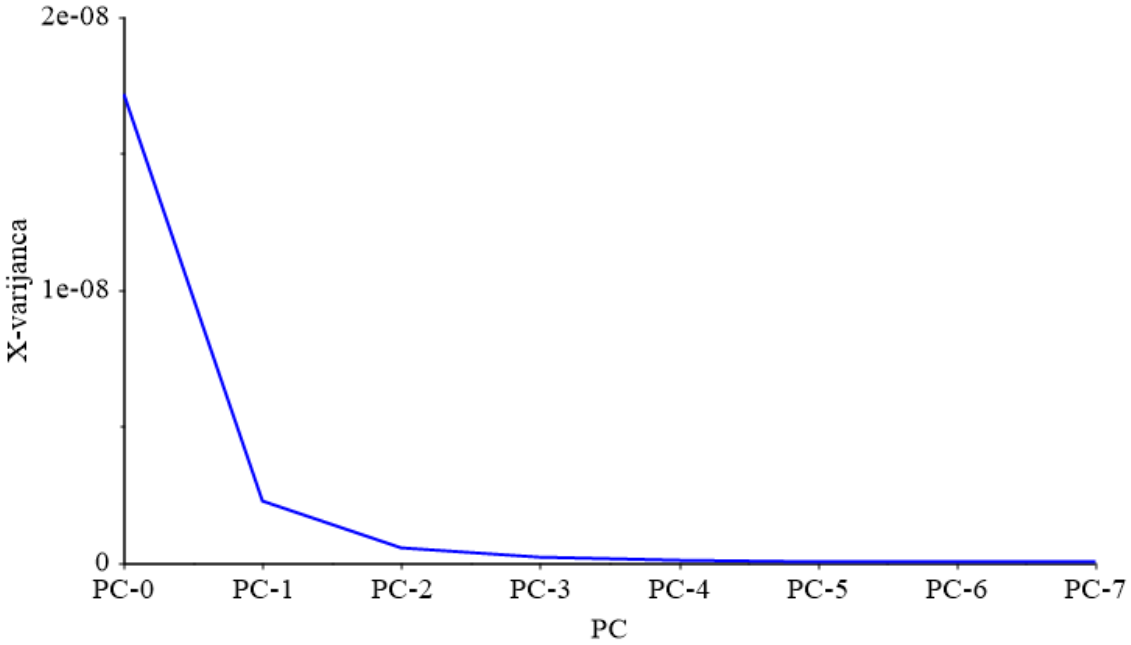
C–H asimetričnog istezanja; spektralno područje $\tilde{\nu}$ = 5300 - 5100 cm^{-1} odgovara O-H kombinacijskim vibracijama; spektralno područje $\tilde{\nu}$ = 4900 - 4500 cm^{-1} proizlazi iz drugog višeg tona C=O istezanja, C-N istezanja i N-H savijanja u ravnini spektralno područje $\tilde{\nu}$ = 4360 - 4450 cm^{-1} odgovara kombinaciji C-H istezanja i CH_2 deformacije regija $\tilde{\nu}$ = 4060 - 4150 cm^{-1} odgovara kombinaciji istezanja C-H, C-C istezanja i C-O-C istezanja (Workman, 2001).

Kako bi se odredilo koliko je PC-ova ključno odnosno optimalno za formiranje PC modela, bilo je potrebno načiniti prikaz kumulativne kalibracijske varijance (Slika 46.).



Slika 46. Kumulativna kalibracijska varijanca za svaki odabrani kumulativni PC.

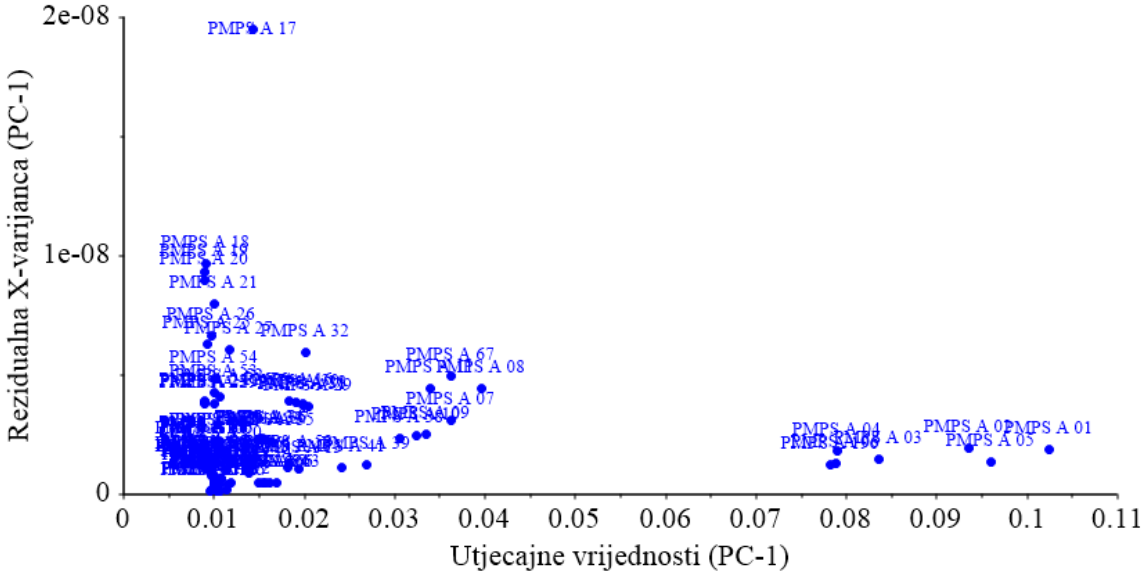
Osim toga, načinjena je i rezidualna kalibracijska varijanca kumulativnih PC-ova (Slika 47.) kako bi se utvrdio optimalan broj PC-ova za model.



Slika 47. Rezidualna kalibracijska varijanca za svaki odabrani kumulativni PC.

Slike 46. i 47. prikazuju koliko svaki PC obuhvaća varijance podataka. Prikaz kumulativne kalibracijske varijance i prikaz rezidualne kalibracijske varijance se koriste kako bi mogli odabrati broj glavnih komponenti (PC-a), koje će se zadržati za PCA model. Ako prvih nekoliko PC-a prikupe većinu informacija (npr. > 80 %), ostatak PC se može zanemariti jer se tako neće izgubiti nikakav ključni podatak iz NIR spektra. Idealna krivulja obiju kumulativnih varijanci bi trebala biti strma u prvom dijelu i zatim se oštro savijati u određenoj točci (PC), nakon koje bi trebala biti skoro vodoravna. Iz Slika 46. i 47. jasno je da bi u ovom slučaju kod PMPS A bio dovoljan jedan ili dva PC-a koja treba uzeti u obzir pri formiranju modela. Uklanjajući PC-ove koji malo doprinose varijanci, projiciramo kompletni skup podataka u manje dimenzionalni prostor, ali zadržavamo većinu ključnih informacija o odabranim NIR spektralnim podacima PMPS A.

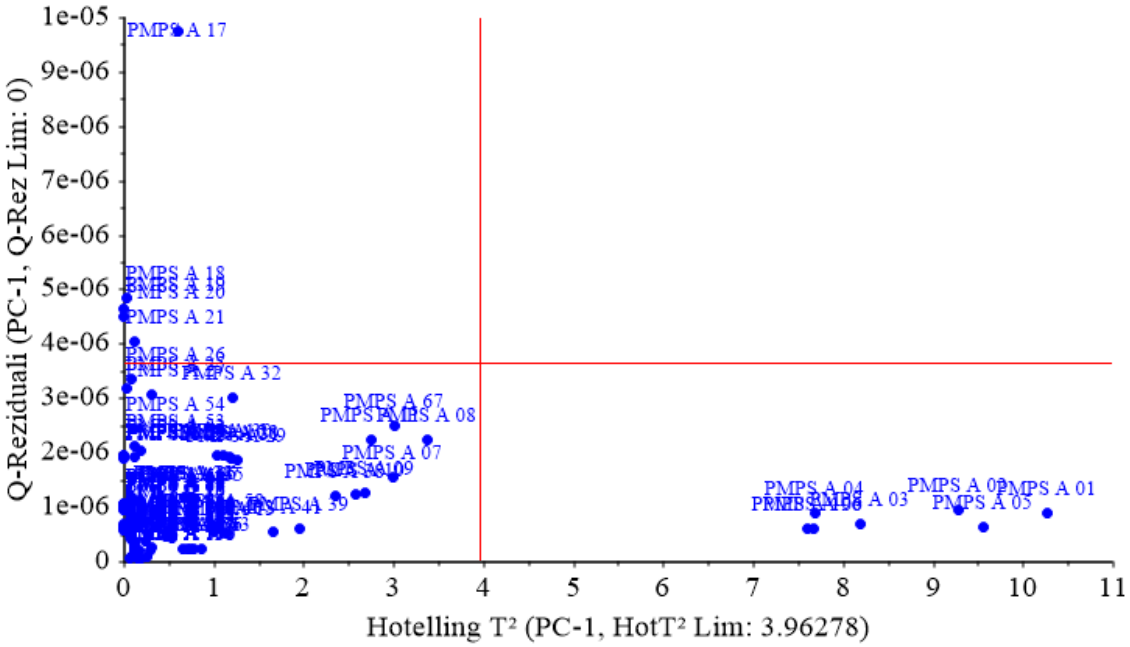
Nakon toga bilo je potrebno odrediti koliki utjecaj pojedini uzorci PMPS A imaju na glavnu komponentu - PC1, te koji su uzorci koji imaju veliku rezidualnu varijancu (Slika 48.).



Slika 48. Rezidualna X-varijanca i utjecajna vrijednost uzoraka PMPS A za PC1.

Na Slici 48. je prikazana rezidualna X-varijanca i utjecajne vrijednosti za PC1. Kao što se može vidjeti na ovoj slici, većina uzoraka se vrlo dobro uklapa u prosječnu kategoriju PMPS A, međutim može se vidjeti i prisutnost potencijalnih netipičnih odnosno ekstremnih uzoraka visokog utjecaja (PMPS A 01-06 i PMPS A 196) i visoke rezidualne X-varijance (PMPS A 17). Kako bi se postigla pravilna klasifikacija odnosno identifikacija nepoznatih uzoraka, koji će se identificirati ovim NIR SIMCA modelom u budućnosti, formiranje ovog modela je bilo potrebno provesti uz korištenje kalibracijskog seta uzoraka PMPS A, koji ne sadrži netipične uzorke. Svaki potencijalni netipični uzorak treba uzeti u obzir pri formiranju ovog modela i utvrditi razloge njihove netipičnosti. Od izuzetne je važnosti razlikovati netipične uzorke od ekstremnih uzoraka. Ekstermni uzorci su prihvatljivi, čak vrlo poželjni, ali bitno različiti proizvodni uzorci po nekoj svojoj karakteristici. Ipak, ekstremne uzorke bi trebalo uzeti u obzir pri formiranju NIR SIMCA modela, ali konačna procjena, bez obzira na statističke rezultate, je subjektivna i odgovornost je na analitičaru da odluči o ovim uzorcima.

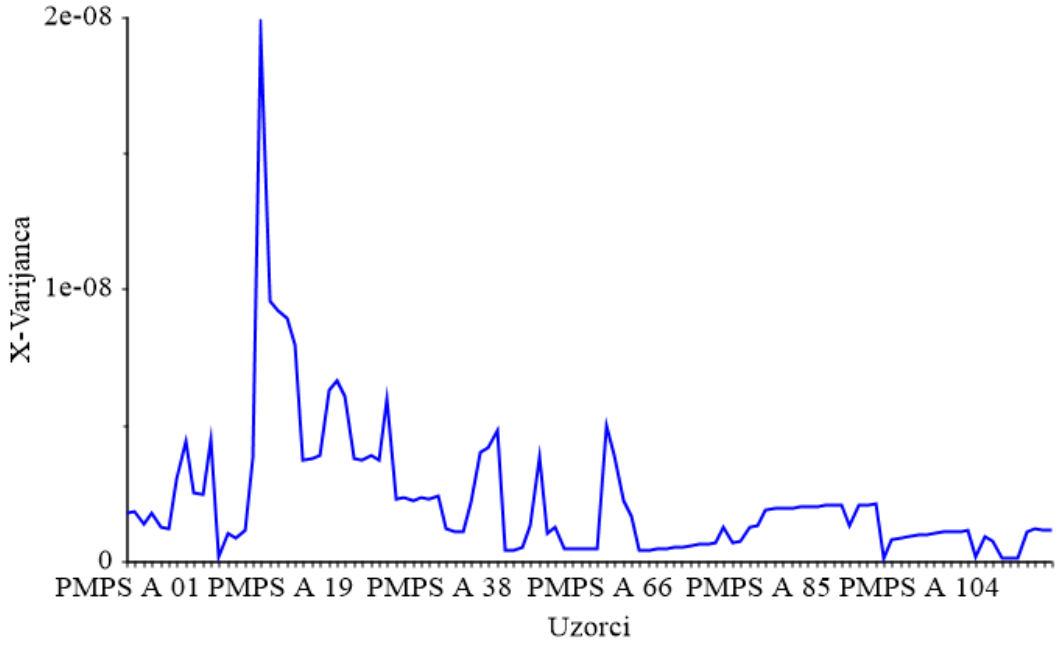
Kako bi utvrdili postojanje netipičnih uzoraka, također je bilo potrebno načiniti i prikaz Hotelling T^2 statistike i Q-reziduala (Slika 49.). Tako su identificirani netipični uzorci.



Slika 49. Hotelling T^2 statistika i Q-reziduali uzoraka PMPS A s pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).

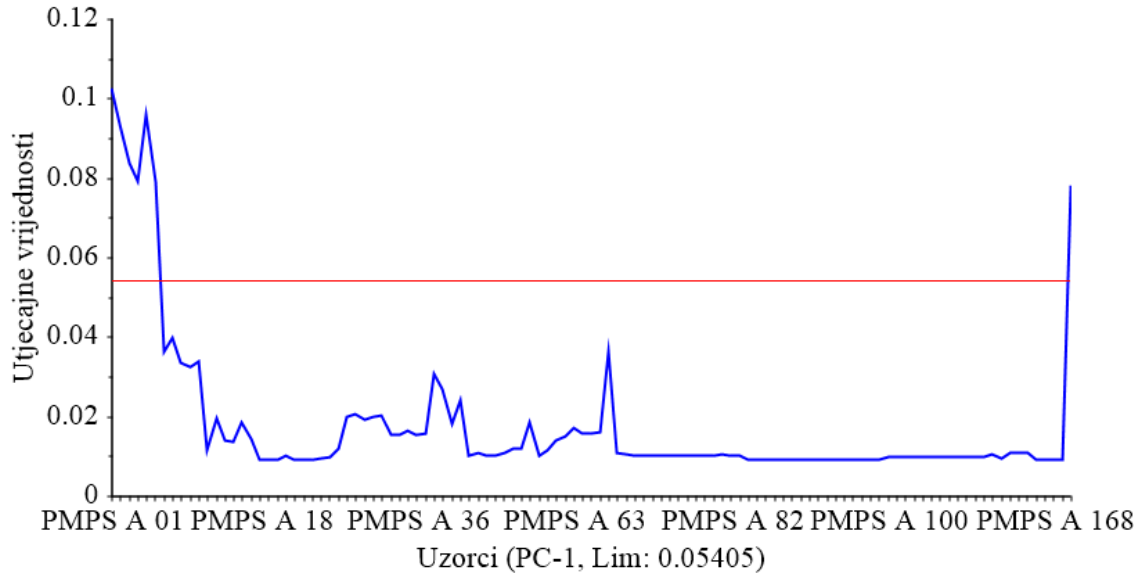
Granična vrijednost za Q reziduale procijenjuje se iz svojstvenih vrijednosti (engl. *eigenvalues*) matrice E (Jackson i Mudholkar,1979). HotellingT^2 kritična granica se temelji na studentovoj t - distribuciji. Uzorci PMPS A 01-06, PMPS A 196 i PMPS A 17 su identificirani kao potencijalni netipični uzorci (Slika 49.). Bez obzira na to, ovi su uzorci uzeti u daljnu statističku analizu, kako bi se utvrdila njihova prikladnost pri formiranju PCA modela. Uzorak PMPS A 17 ima visoku vrijednost Q reziduala, odnosno veliku udaljenost do modela, što ukazuje na to da uzorak nije u skladu s modelom. Potrebno ga je dalje statistički obraditi kako bi se utvrdio uzrok visoke rezidualne varijance ovog uzorka. Uzorci PMPS A 01-06, PMPS A 196 imaju veliki utjecaj na model i od izuzetne je važnosti utvrditi dali ovi uzorci utječu na većinu varijance pojedine glavne komponente te na taj način utječu i na prediktivne sposobnosti nepoznatih uzoraka budućim formiranim modelom.

Kod svih uzoraka bilo je potrebno načiniti analizu rezidualne X-varijance pojedinačnih uzoraka (Slika 50.), kako bi se sa još većom sigurnošću identificirali netipični uzorci.

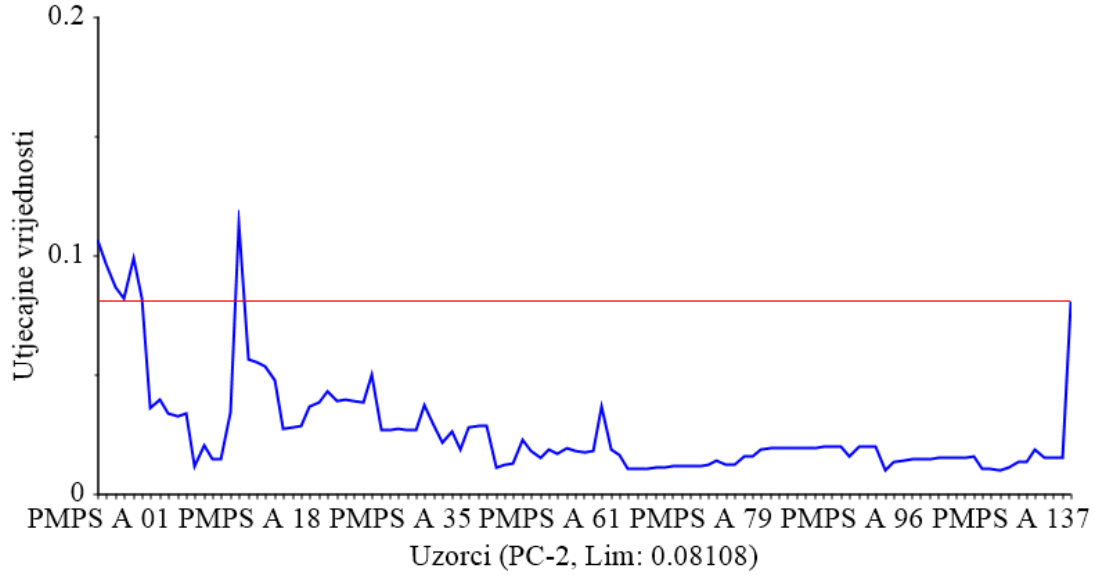


Slika 50. Rezidualna X-varijanca uzoraka PMPS A.

Iz Slike 50. jasno se vidi koliko preostala X-varijanca pojedinih uzorka varira u usporedbi s drugim uzorcima PMPS A. Ovdje je potvrđeno da prethodno identificirani netipični uzorak, PMPS A 17 (Slika 49.) ima veću X-varijance u usporedbi s preostalim PMPS A uzorcima. Nadalje, načinjena je i analiza utjecaja svakog PMPS A uzorka na PC1 i PC2 (Slike 51. i 52.).



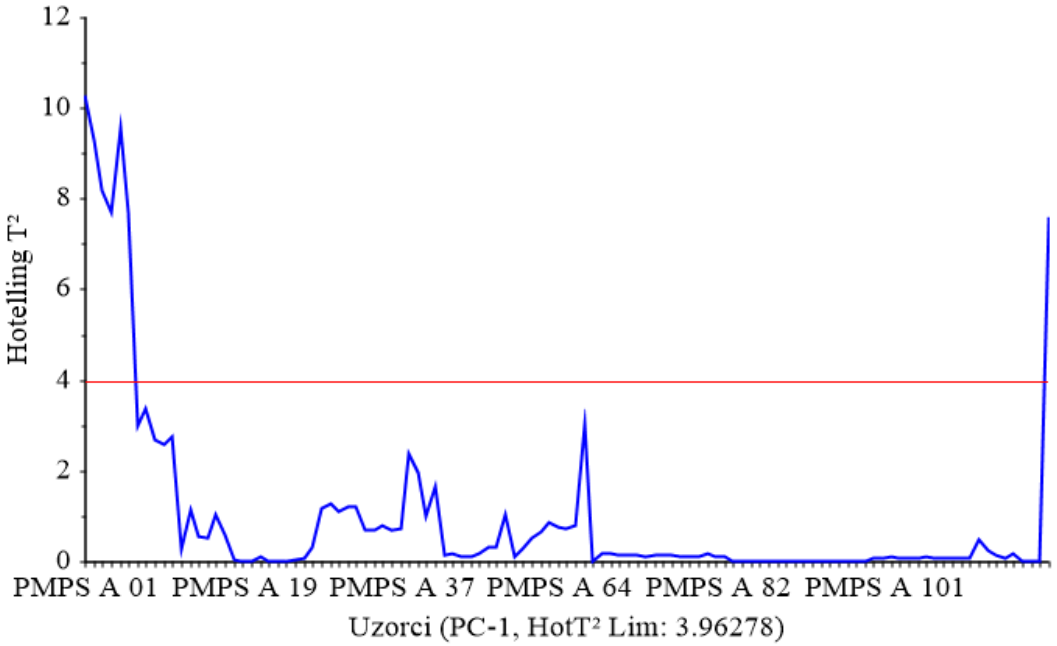
Slika 51. Utjecajne vrijednosti uzoraka PMPS A za PC1 sa pripadajućom kritičnom vrijednosti (crvena linija).



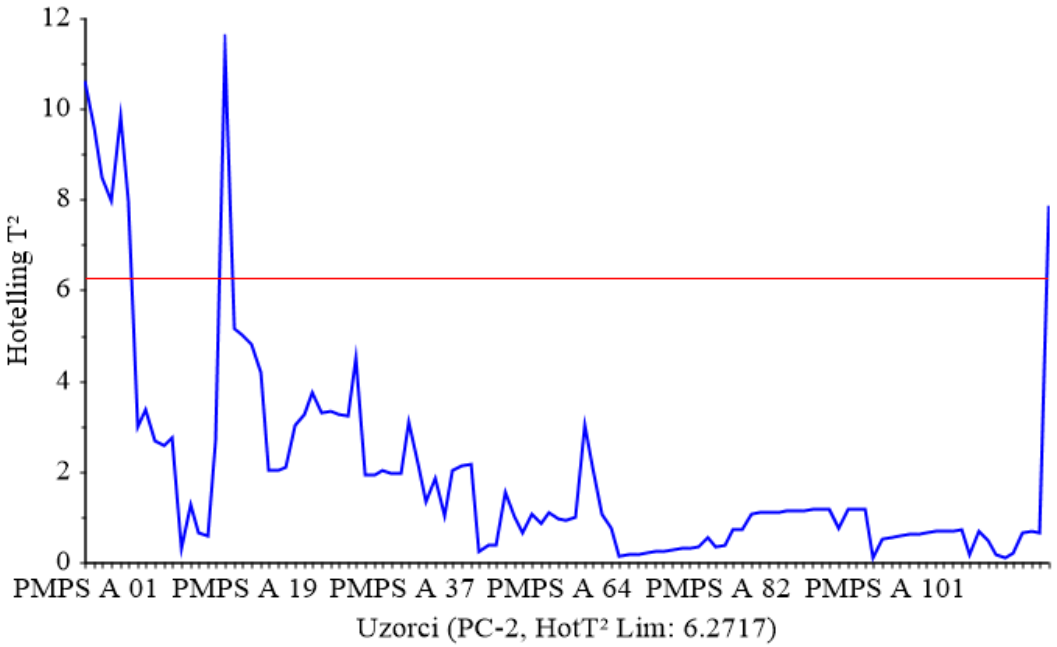
Slika 52. Utjecajne vrijednosti uzoraka PMPS A za PC2 sa pripadajućom kritičnom vrijednosti (crvena linija).

Uzorci sa visokim utjecajem na PC1 i PC2 se razlikuju od preostalih prosječnih uzoraka PMPS A tj. velika je vjerojatnost da su ovi uzorci PMPS A netipični uzorci. Najveći utjecaj na model, u odnosu na ostale uzorke kalibracijskog seta, kako je vidljivo iz Slike 51 i Slike 52, imaju uzorci PMPS A 01 i PMPS A 05, te PMPS A 17.

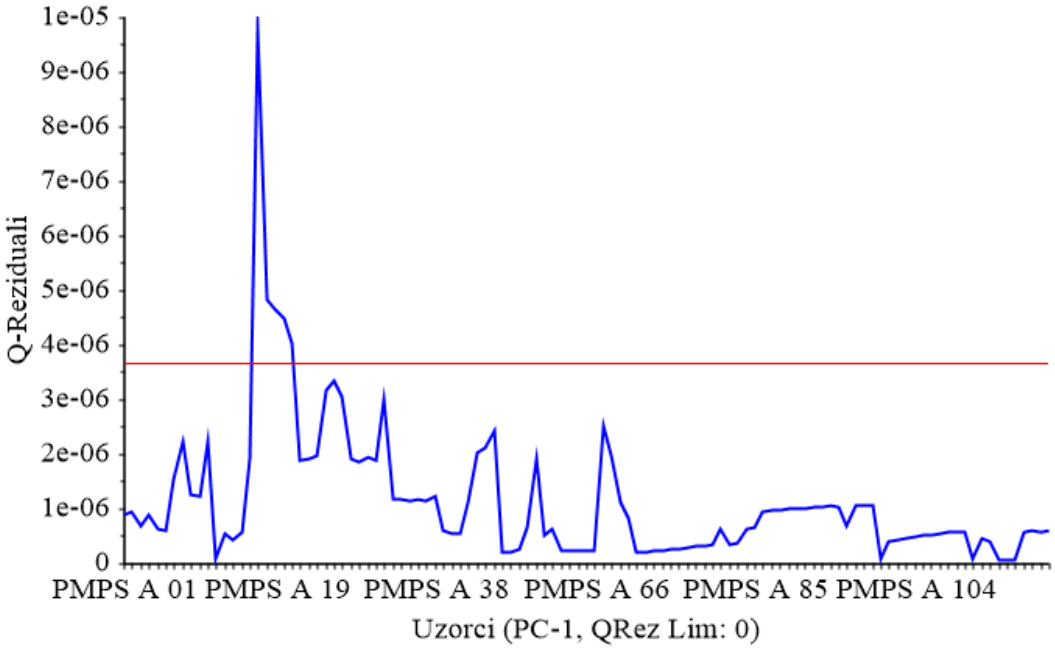
Osim toga, načinjena je Hotelling T^2 statistika (Slike 53. - 56.), kao mjera varijacije svakog uzorka unutar PCA modela., odnosno radi provjere varijabilnosti projiciranih podataka u novom prostoru glavnih komponenti. Također su načinjeni Q reziduali kao mjera razlike između uzorka i njegove projekcije u model, te su tako identificirali potencijalni netipični uzorci sa znanto većom sigurnošću.



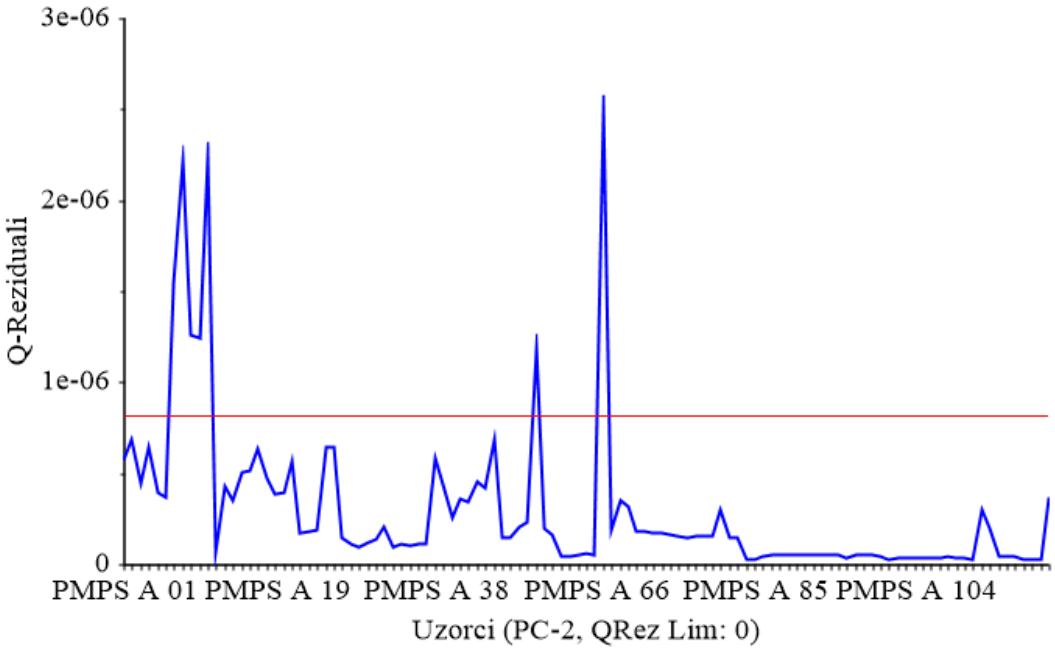
Slika 53. Hotelling T^2 statistika uzoraka PMPS A za PC1 sa pripadajućom kritičnom vrijednosti (crvena linija).



Slika 54. Hotelling T^2 statistika uzoraka PMPS A za PC2 sa pripadajućom kritičnom vrijednosti (crvena linija).



Slika 55. Q reziduali uzoraka PMPS A za PC1 s pripadajućom graničnom linijom (crvena linija).

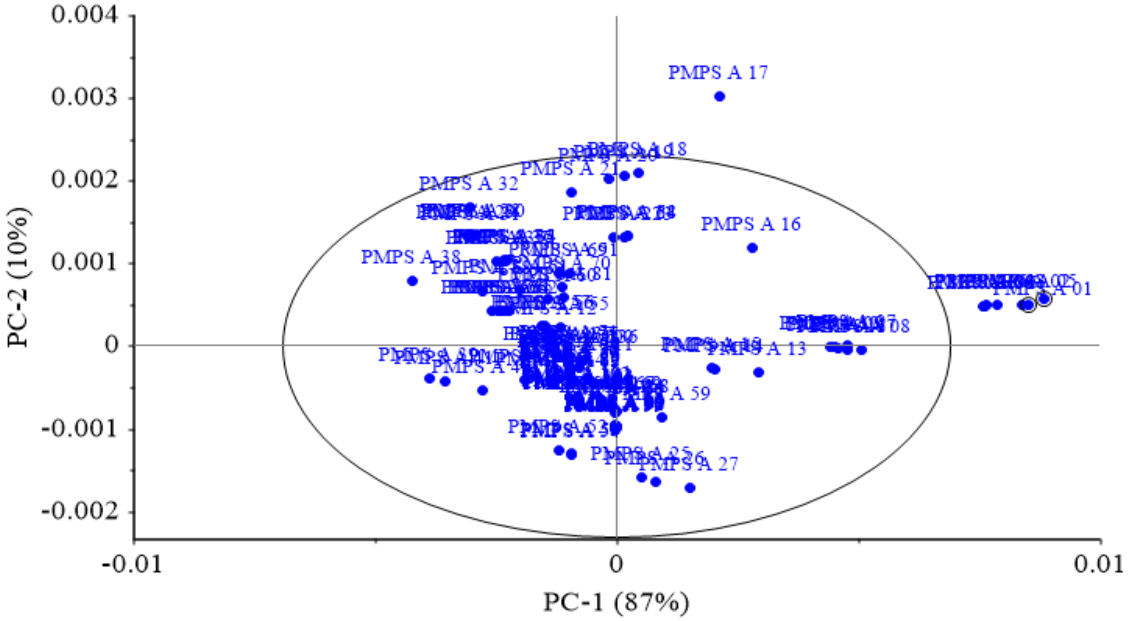


Slika 56. Q reziduali uzoraka PMPS A za PC2 s pripadajućom graničnom linijom (crvena linija).

Hotelling T^2 statistika (Slike 53 - 54.) prikazuje koliko dobro formirani PCA model opisuje odabrani uzorak PMPS A, odnosno Hotelling T^2 vrijednosti opisuju mjeru varijacije svakog uzorka unutar modela. Hotelling T^2 statistika je u linearnom odnosu sa utjecajnom vrijednosti uzoraka. Q-reziduali opisuju udaljenost uzorka od modela, odnosno predstavljaju veličinu varijacije koja zaostaje u svakom uzorku nakon projekcije kroz model. (Slike 55 - 56.)

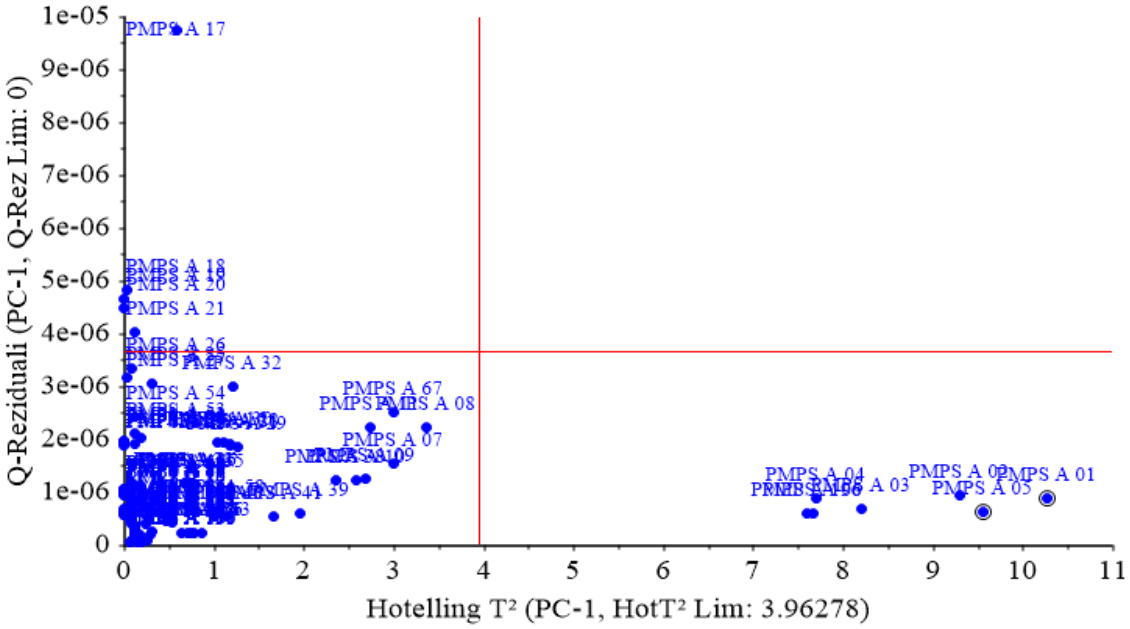
Sukladno rezultatima prikazanim na Slikama 48. - 56. nedvojbeno su identificirani netipični uzorci PMPS A 01 i PMPS A 05.

Kako ekstremni uzorci imaju važnu ulogu u statističkoj analizi NIR spektralnih podataka, izuzetno su važna karakteristika bilo kojeg skupa podataka i ne smije ih se zamijeniti s netipičnim uzorcima. Statističkom analizom eksperimentalno dobivenih NIR spektralnih podataka za PMPS A, utvrđeno je da svakako treba izdvojiti uzorke PMPS A 01 i PMPS A 05 kao netipične uzorke. Ovi uzorci imaju veliki utjecaj na modeli i njihovim se uklanjajem mijenja izvedba modela. Utjecaj ovih uzoraka je jedinstven i važan, ali zbog nemogućnosti da se prikupi veća količina ovakvih uzoraka kako bi se mogao stabilizirati model, uzorke je potrebno izdvojiti iz kalibracijskog seta uzoraka. Vizualnim pregledom preostalih ekstremnih uzoraka, te pregledom sirovih podataka ustanovljeno je da su uzorci različite gustoće pakiranja, od ostalih uzoraka trening skupa, te je velika rezidualna varijanca posljedica dijela spektra nevažnog za identifikaciju PMPS. Slijedom toga, preostali uzorci identificirani su kao ekstremni uzorci poželjni za formiranje modela te se neće izdvojiti kao netipični uzorci već će se zadržati u kalibracijskom skupu uzoraka. Zbog toga je bilo potrebno nanovo načiniti PCA model preostalih NIR spektralnih podataka PMPS A, ali ovaj put bez netipičnih uzoraka. U ponovljenoj statističkoj obradi označeni su netipični uzorci (Slika 57.). Ova se ponovljena statistička obrada može usporediti sa podacima na Slici 39.



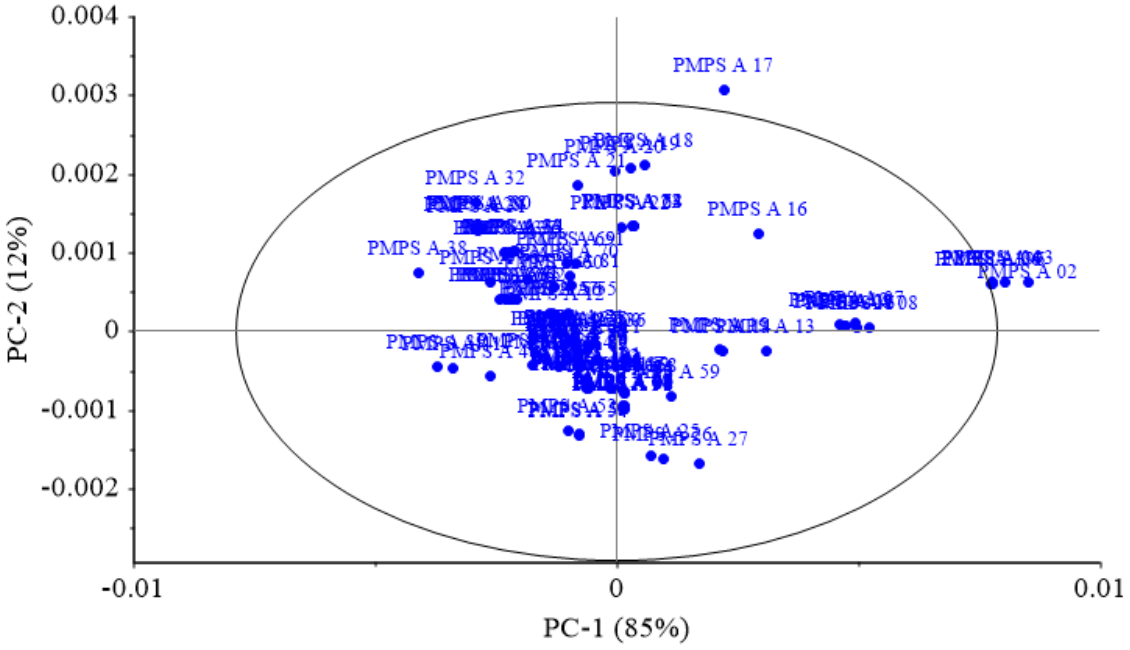
Slika 57. Raspodjela faktorskih bodova PC1 i PC2 matematički obrađenih (Savitzky-Golay glačanje 3.9 s drugom derivacijom) NIR spektara PMPS A u spektralnom području $\tilde{\nu}$ = 7768 - 3695 cm^{-1}. Elipsa označava 95 %-tni interval pouzanosti (Hotelling T^2 statistika).

Netipični uzorci istaknuti i na Slici 58. (ovdje ispod).



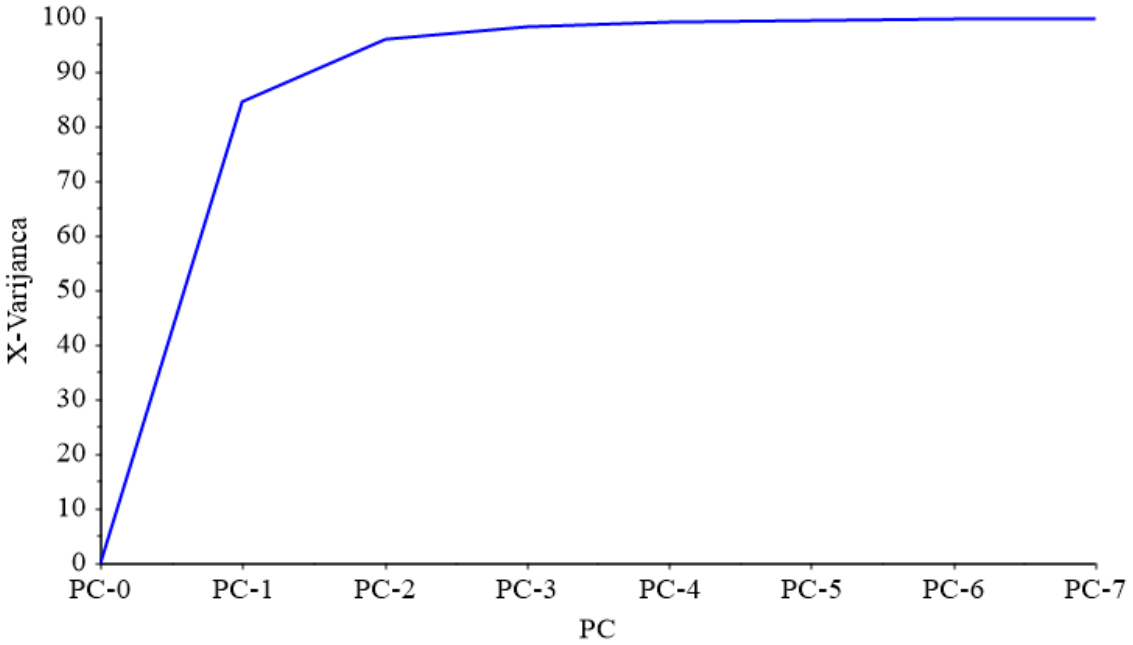
Slika 58. Hotelling T^2 statistika i Q-reziduali uzoraka PMPS A za PC1 s označenim netipičnim uzorcima i pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).

Ponovljena PCA stataistička analiza NIR spektralnih podataka PMPS A nakon isključivanja netipičnih uzoraka prikazana je na donjoj slici.



Slika 59. Raspodjela faktorskih bodova PC1 i PC2 matematički obrađenih (Savitzky-Golay glačanje 3.9 s drugom derivacijom) NIR spektara PMPS A u području $\tilde{\nu}$ = 7768 - 3695 cm^{-1}. Elipsa označava 95 %-tni interval pouzanosti (Hotelling T^2 statistika).

Nakon načinjene ponovne statističke PCA analize NIR spektralnih podataka za PMPS A, napravljen je i prikaz kumulativne kalibracijske varijance (Slika 60. i Tablica 1.) kako bi se utvrdio optimalni broj PC-ova za NIR PCA model za PMPS A.



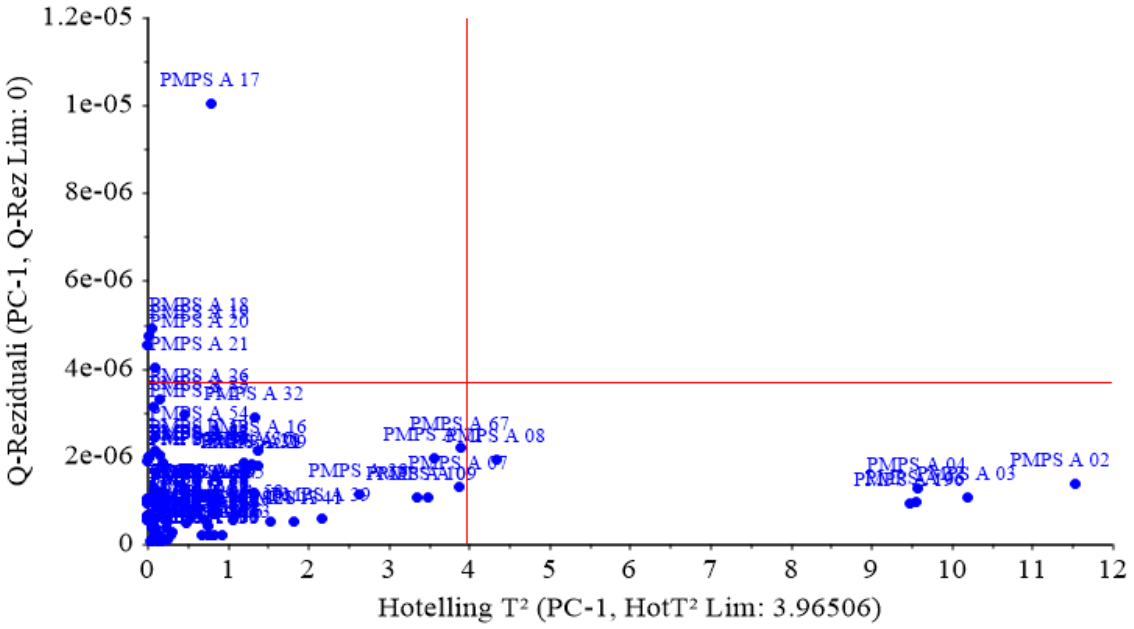
Slika 60. Kumulativna kalibracijska varijanca za svaki PC nakon uklanjanja netipičnih uzoraka.

Tablica 1. Kumulativna kalibracijska varijanca za svaki odabrani kumulativni PC nakon uklanjanja netipičnih uzoraka.

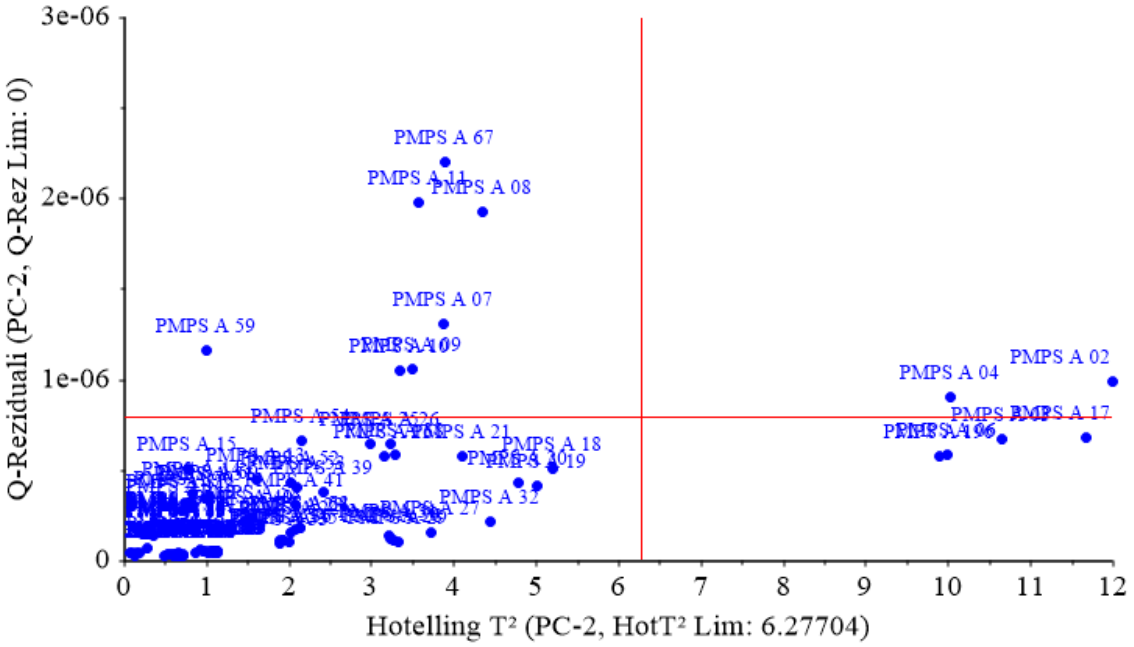
	PC0	PC1	PC2	PC3	PC4	PC5	PC6
Kalibracija	0	84,5416	96,0687	98,2532	99,1540	99,4610	99,6590

Iz Tablice 1. se može vidjeti da jedan PC (PC1) obuhvaća 85 % ukupne varijance,dva PC-a (PC1 i PC2) obuhvaćaju oko 96 % ukupne varijance, dok preostali PC-ovi obuhvaćaju neznatno više ukupne varijance i može se smatrati da obuhvaćaju šumove, koji nisu potrebni za formiranje PCA modela. Iz dobivenih rezultata odabrani broj PC, PCA modela za PMPS A je dva.

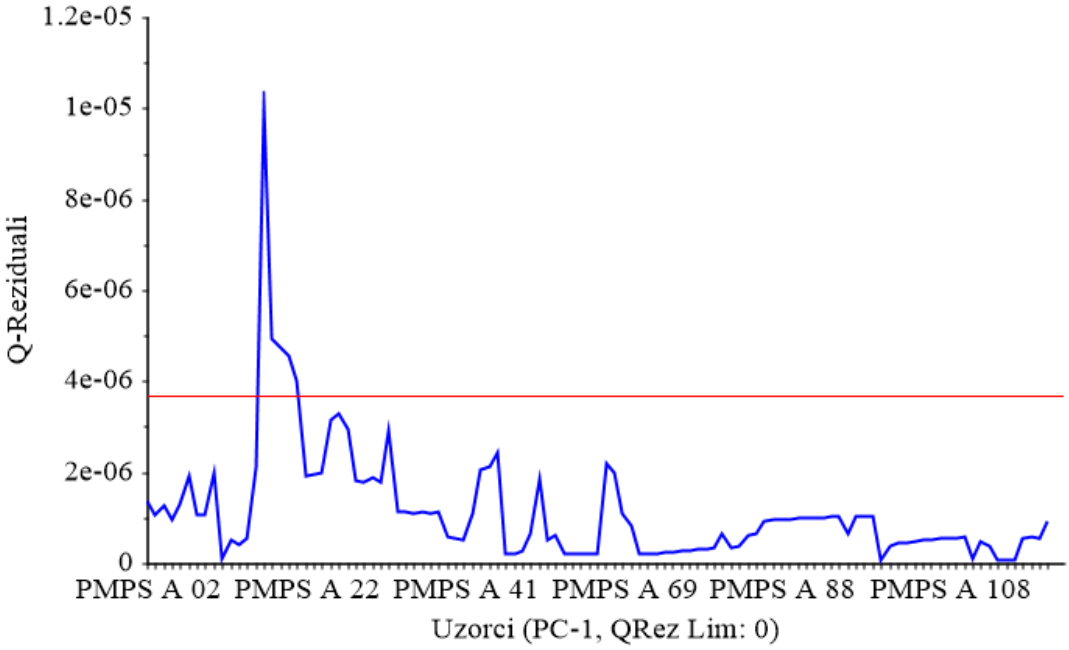
Također, načinio se: prikaz utjecajnih vrijednosti uzoraka, Hotelling T^2 statistika i Q-reziduali u cilju identificiranja netipičnih PMPS A uzoraka.



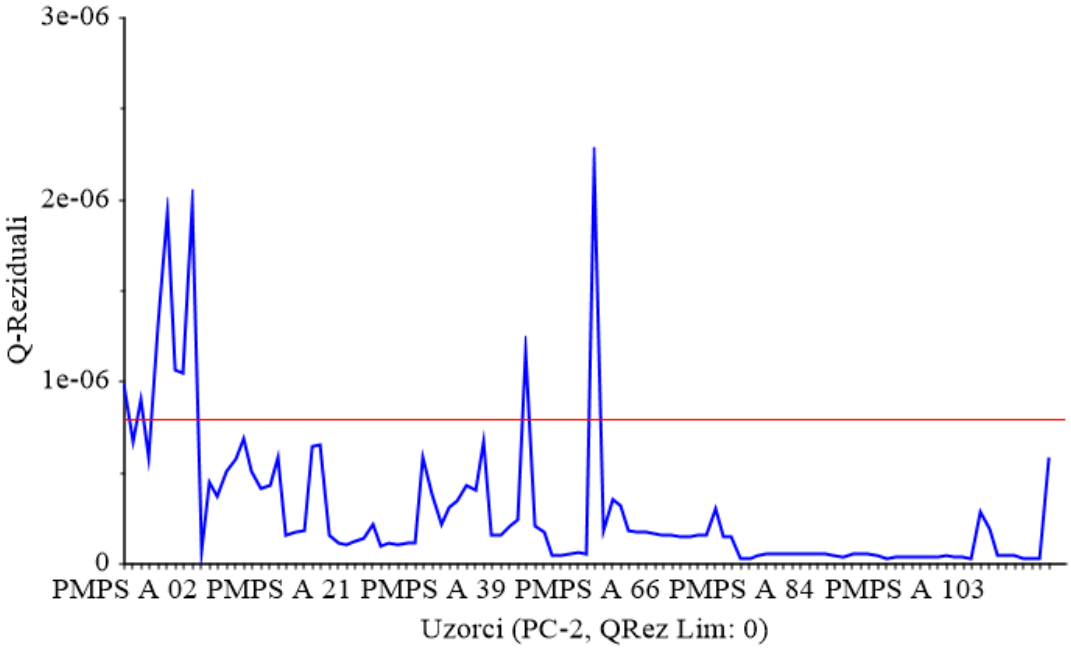
Slika 61. Hotelling T^2 statistika i Q-reziduali uzoraka PMPS A za PC1 s pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).



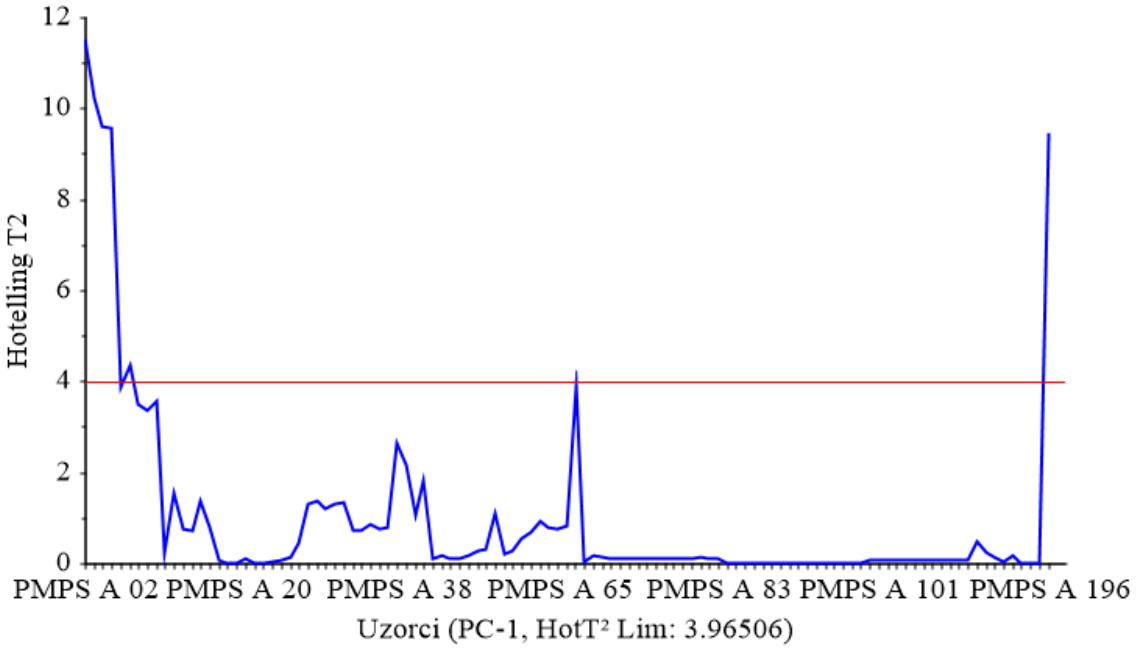
Slika 62. Hotelling T^2 statistika i Q-reziduali uzoraka PMPS A za PC2 s pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).



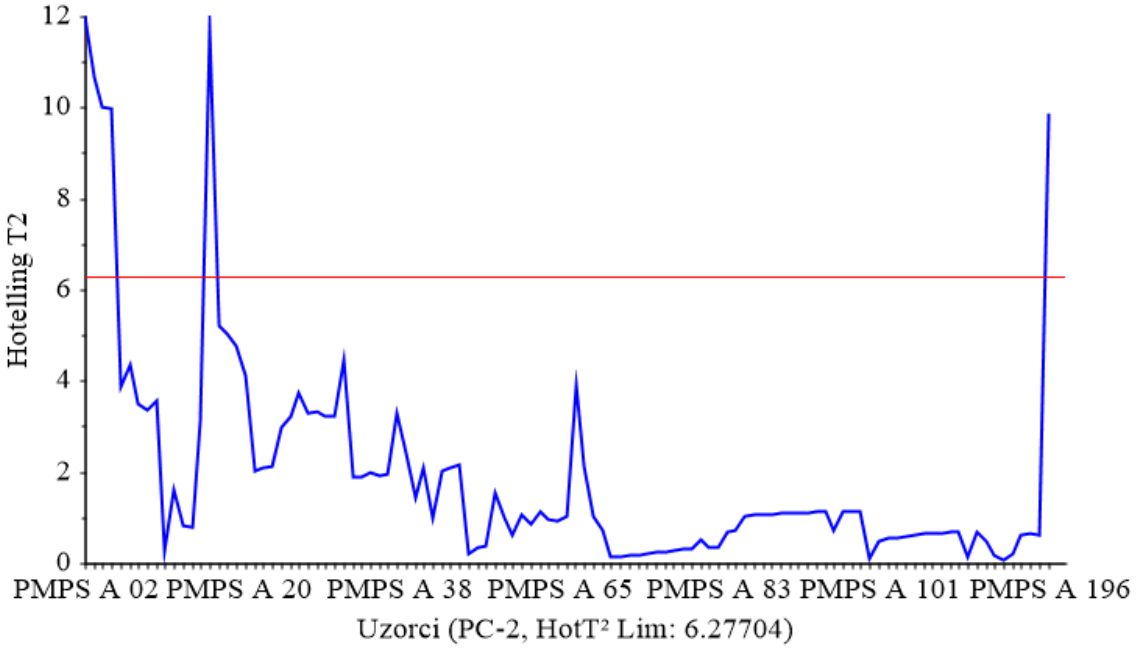
Slika 63. Q reziduali uzoraka PMPS A za PC1 s pripadajućom graničnom linijom (crvena linija).



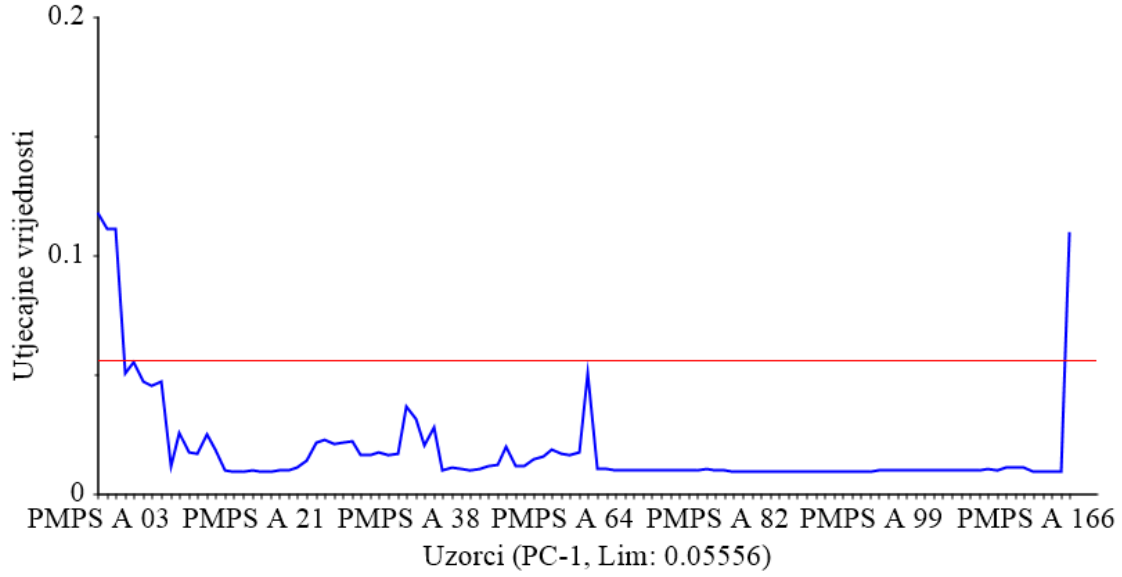
Slika 64. Q reziduali uzoraka PMPS A za PC2 s pripadajućom graničnom linijom (crvena linija).



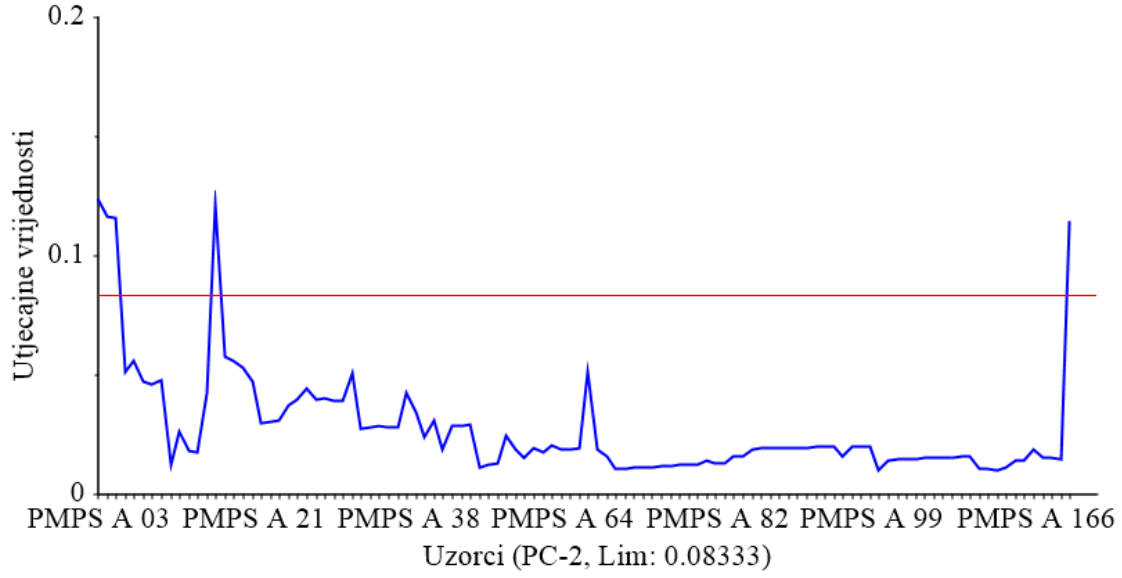
Slika 65. Hotelling T^2 statistika uzoraka PMPS A za PC1 s pripadajućom kritičnom vrijednosti (crvena linija).



Slika 66. Hotelling T^2 statistika uzoraka PMPS A za PC2 s pripadajućom kritičnom vrijednosti (crvena linija).



Slika 67. Utjecajne vrijednosti uzoraka PMPS A za PC1 sa pripadajućom kritičnom vrijednosti (crvena linija).



Slika 68. Utjecajne vrijednosti uzoraka PMPS A za PC2 sa pripadajućom kritičnom vrijednosti (crvena linija).

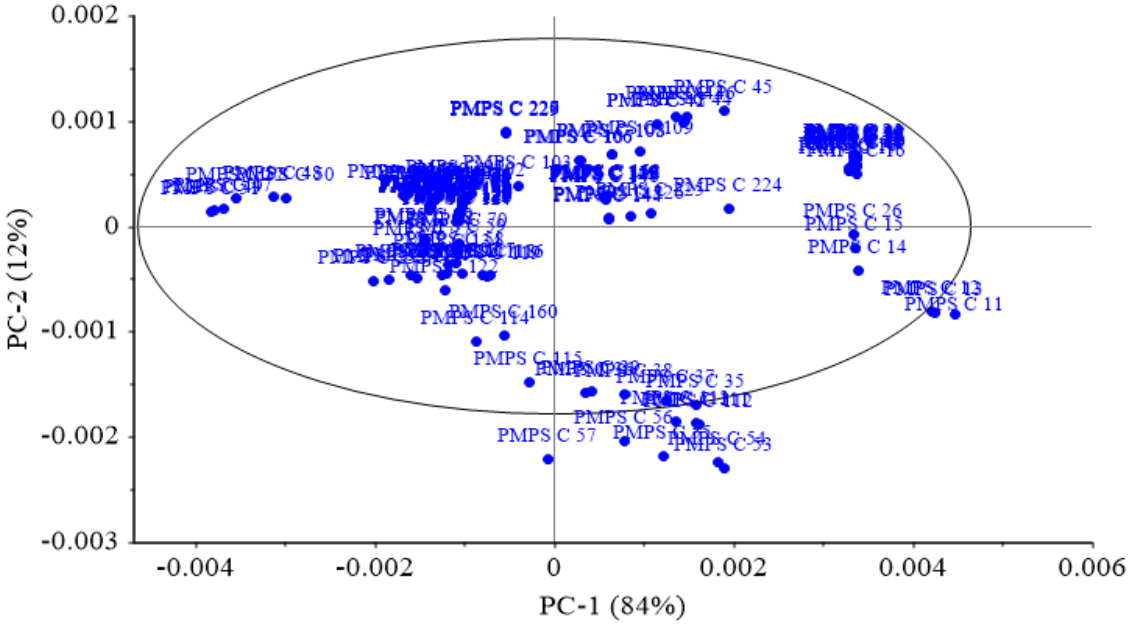
Hotelling T^2 statistika, Q-reziduali te utjecajne vrijednosti uzoraka na glavne komponente (Slike 61. - 68.) daju kompletan uvid na prisutnost oba tipa netipičnih uzoraka, prikazuju koliko

dobro formirani PCA model opisuje uzorke kalibracijskog seta uzoraka, rezidualne varijance, odnosno profil uzoraka kalibracijskog skupa PMPS A nakon izuzimanja netipičnih uzoraka. Može se vidjeti prisutnost uzoraka viših Hotelling T^2 vrijednosti, i uzoraka viših vrijednosti Q reziduala. Ovi ekstremni uzorci od ključne su nam važnosti za formiranje robustnog i visoko učinkovitog modela, obzirom da je vizualnim pregledom utvrđeno da su ovi uzorci različitih fizičkih karakteristika (gustoće pakiranja te površina uzorka) koje predstavljaju realne proizvodne slučajeve i štoviše, poželjno je da su uključeni u formiranje modela. Iz dobivenih rezultata možemo zaključiti da nema novo identificiranih netipičnih uzoraka koje je potrebno izuzeti iz kalibracijskog skupa PMPS A.

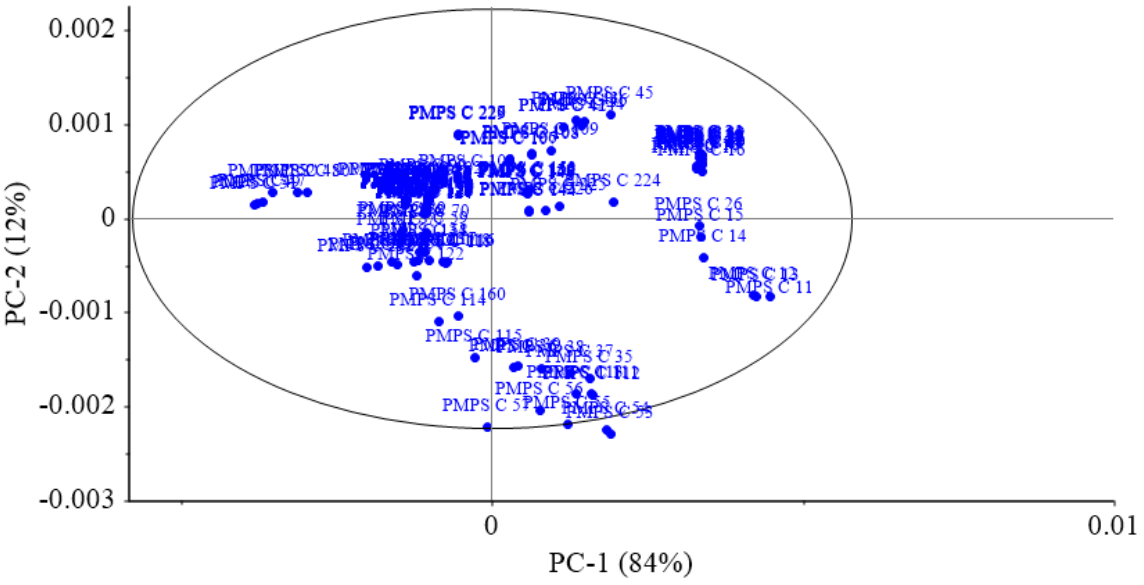
4.2.5.1.2. Analiza glavnih komponenti NIR spektralnih podataka za PMPS C

PCA je provedena na NIR spektralnim podacima iz kalibracijskog skupa uzoraka PMPS C, koji uključuje 153 NIR spektara. Ostali NIR spektralni podaci PMPS C (uzorci iz test seta1 i vanjskog test seta) nisu korišteni u ovoj analizi, već su se koristili isključivo za optimizaciju i validaciju modela.

Kako je to načinjeno i za PMPS A (vidi poglavlje 4.2.5.1.1.), i ovdje se u okviru PCA načinilo: raspodjela faktorskih bodova, opterećenja, zatim prikaz utjecajnih vrijednosti za PC1 i PC2, Hotelling T^2 statistika i Q-reziduali u cilju identificiranja netipičnih PMPS C uzoraka (ovdje ispod).

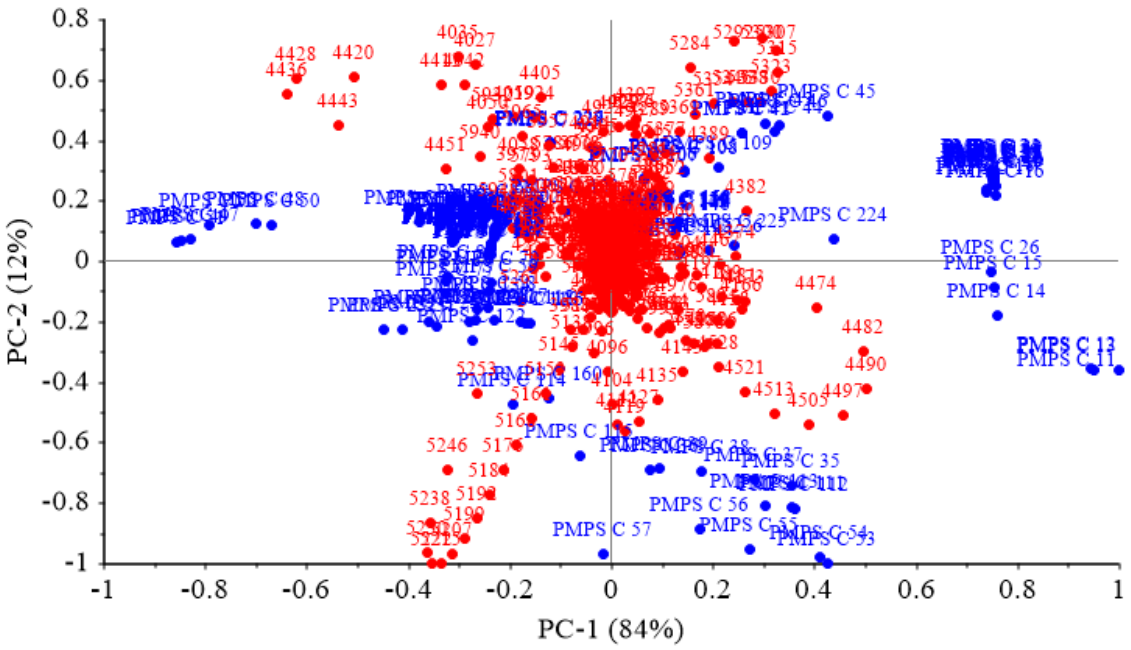


Slika 69. Raspodjela faktorskih bodova PC1 i PC2 matematički obrađenih (Savitzky-Golay glačanje 3.9 s drugom derivacijom) NIR spektara PMPS C u spektralnom području $\tilde{\nu}$ = 7768 - 3695 cm^{-1}. Hotelling T^2 elipsa označava 95 %-tni interval pouzanosti sa centrom modela u sjecištu dvaju pravaca (sivo).



Slika 70. Raspodjela faktorskih bodova PC1 i PC2 matematički obrađenih (Savitzky-Golay glačanje 3.9 s drugom derivacijom) NIR spektara PMPS C u spektralnom području $\tilde{\nu}$ = 7768 - 3695 cm^{-1}. Hotelling T^2 elipsa označava 99 %-tni interval pouzanosti sa centrom modela u sjecištu dvaju pravaca (sivo).

Slike 69. i 70. daju informaciju o grupiranju uzoraka PMPS C, trendu među ovim uzorcima te netipičnim vrijednostima. Na ovim slikama su prikazani faktorski bodovi prve i druge glavne komponente (PC1 i PC2). Dijagram faktorskih bodova PC1 i PC2 je posebno koristan, jer ove dvije komponente sažimaju više varijacija u NIR spektralnim podacima nego bilo koji drugi par komponenti. Hotelling T^2 elipsa (interval pouzdanosti 95 % i 99%, Slike 69-70.) je dobar način za detekciju netipičnih uzoraka tj. vrijednosti ili ekstremnih uzoraka, koje je svakako potrebno dalje analizirati. Prvi faktor, odnosno prva glavna komponenta (PC1) oduhvaća 84 % varijance, dok druga glavna komponenta (PC2) obuhvaća 12 % varijance. Slike 69 i 70. prikazuju jednoliko raspodjeljene uzorke kroz cijelo područje. Uzorci PMPS C 53 i PMPS C 54 izlaze izvan Hotelling T^2 elipse (99 % pouzdanosti) i predstavljaju moguće netipične uzorke te ih je potrebno dodatno analizirati statističkim metodama (ovdje ispod).

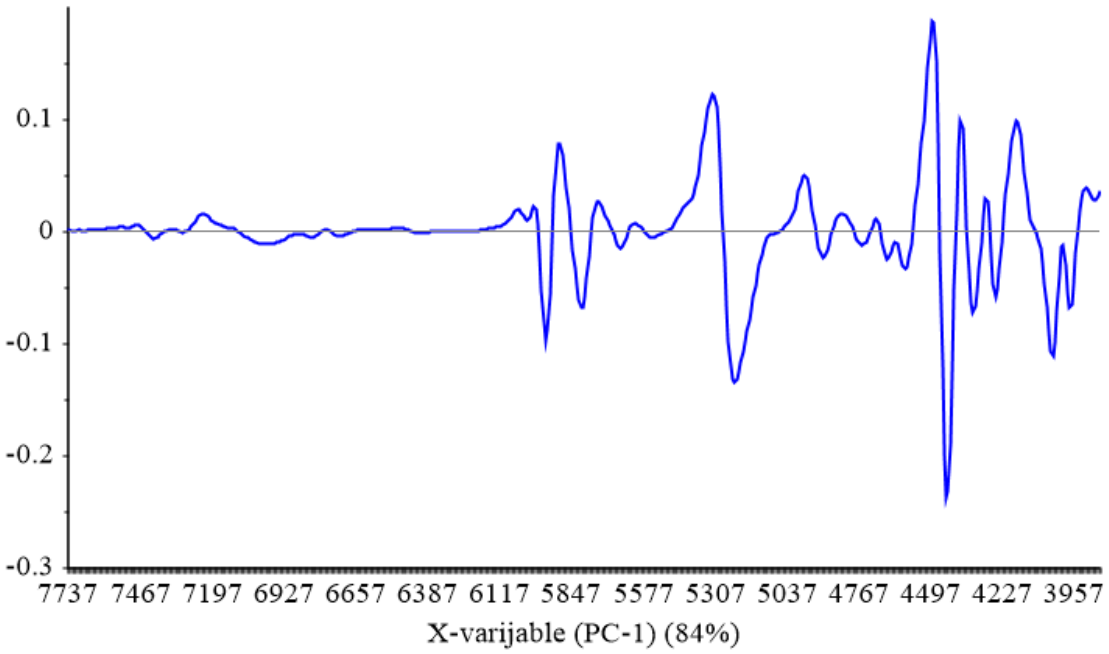


Slika 71. Raspodjela faktorskih bodova PC1 i PC2 matematički obrađenih (Savitzky-Golay glačanje 3.9 s drugom derivacijom) NIR spektara PMPS C u području $\tilde{\nu}$ = 7768 - 3695 cm^{-1} zajedno sa PC1 i PC2 opterećenjem.

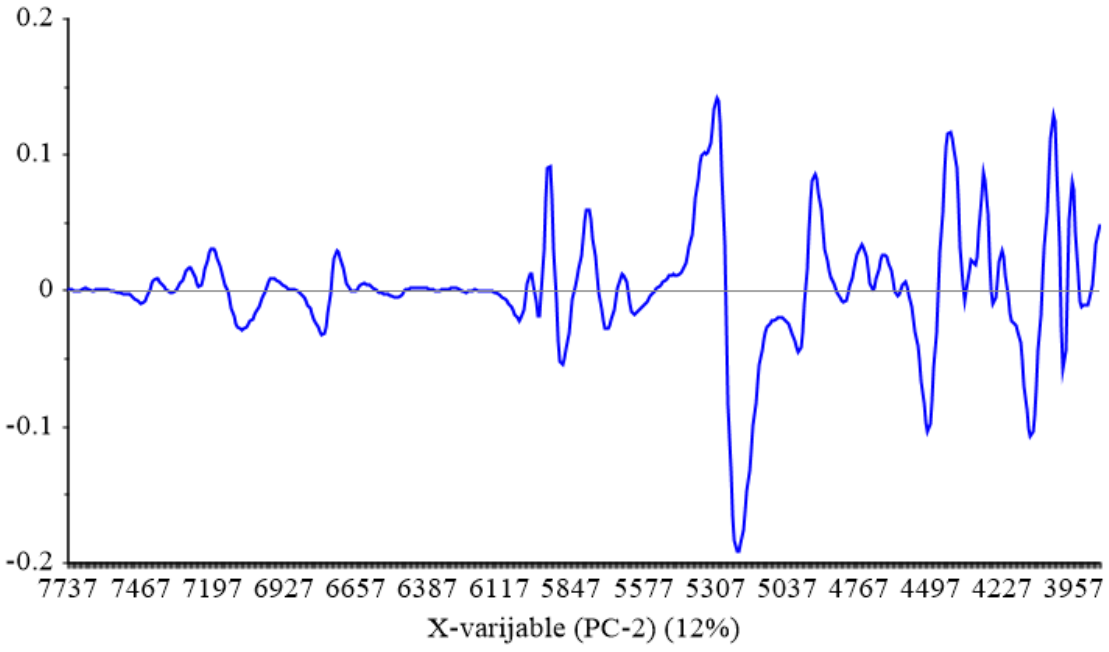
Na Slici 71. može se vidjeti koje varijable su odgovorne za razliku među ovim uzorcima. Opterećenja opisuju set podataka u smislu doprinosa varijabli i njihove međusobne korelacije. Svaka glavna komponenta nosi svoja opterećenja za cijeli set varijabli. Geometrijski gledano,

faktorsko opterećenje je kosinus kuta između varijabli i glavne komponente. Što je taj kut manji, to će opterećenje biti veće i obrnuto.

Prikaz faktorskih bodova i opterećenja (Slike 70. i 71.) je dobar način za analizu odnosa između varijabli i identifikaciju najutjecajnijih varijabli u formiranju PCA modela. Najvažniji je međuodnos PC1 i PC2, budući da predstavlja najveću varijancu u podacima. Varijable koje su međusobno blizu (Slika 71.) utječu na PCA model na slične načine, što je također pokazatelj da su u uzajamnoj vezi. U slučaju kada varijable leže na PC1 i PC2, kosinus kuta između PC1 i varijable je jednak 1, što znači da su varijable u potpunosti opisane, redom, prvom (PC1) odnosno drugom glavnom komponentom (PC2). Varijable u dijagonalno suprotnim kvadrantima imaju tendenciju negativne korelacije. Varijable blizu centra (Slika 71.) se u ovom dvodimenzionalnom prikazu za ova dva PC-a ne mogu interpretirati tj. varijable na sjecištu nisu dobro opisane ni s jednom od dviju glavnih komponenti. Također, važna je i udaljenost točaka od centra (Slika 71.): što je varijabla dalje od centra, to ima više utjecaja u formiranju PCA modela. Uzorci PMPS C koji se nalaze na vektoru glavne komponente u istom smjeru kao i dane varijable, imaju veliku vrijednost (utjecaj) za tu varijablu. Uzorci koji se nalaze na vektoru glavne komponente u suprotnom smjeru, imaju nisku vrijednost za tu varijablu. Dakle, ako varijabla ima veliko pozitivno ili negativno opterećenje, to znači da je varijabla važna za dotičnu glavnu komponentu. Kod formiranja NIR modela, linijska opterećenja predstavljaju izvrsnu varijantu interpretacije opterećenja NIR spektara, jer linijska opterećenja imaju profil sličan originalnim podacima. Zbog toga su načinjene Slike 72. i 73.

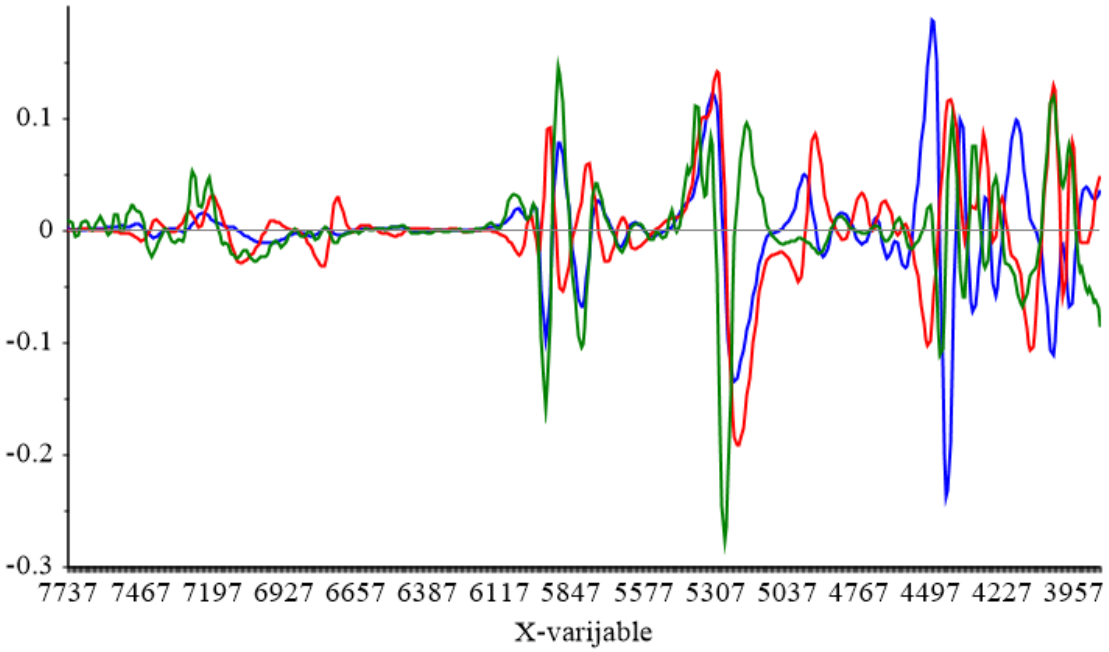


Slika 72. PC1 opterećenje po valnim brojevima ($\tilde{\nu}$) dobivena PCA analizom NIR spektara PMPS C.



Slika 73. PC2 opterećenje po valnim brojevima ($\tilde{\nu}$) dobivena PCA analizom NIR spektara PMPS C.

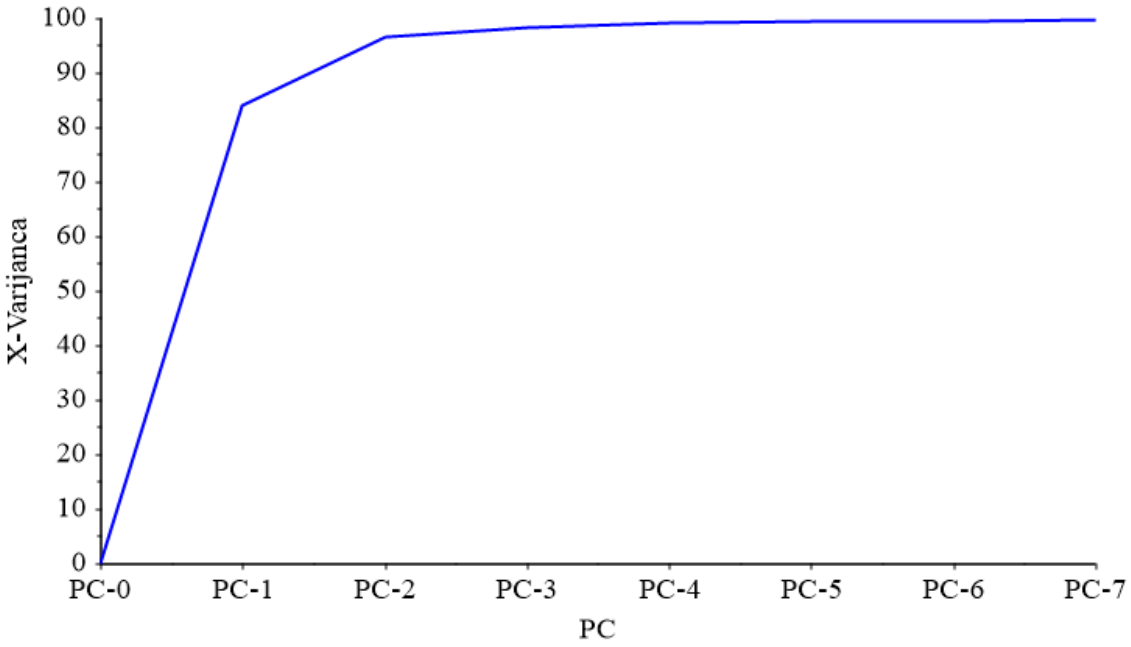
Valni brojevi ($\tilde{\nu}$) s najvećim opterećenjima najviše doprinose pojedinoj komponenti (PC1 i PC2). Dodatno, preklopljena su opterećenja za PC1, PC2, PC3, kako bi se utvrdile najvažnije spektralne regije, koje definiraju glavne komponente (Slika 74.).



Slika 74. PC1 (plava), PC2 (crvena), PC3 (zelena) opterećenja po valnim brojevima dobivenih PCA analizom za skup od 153 NIR spektara PMPS C.

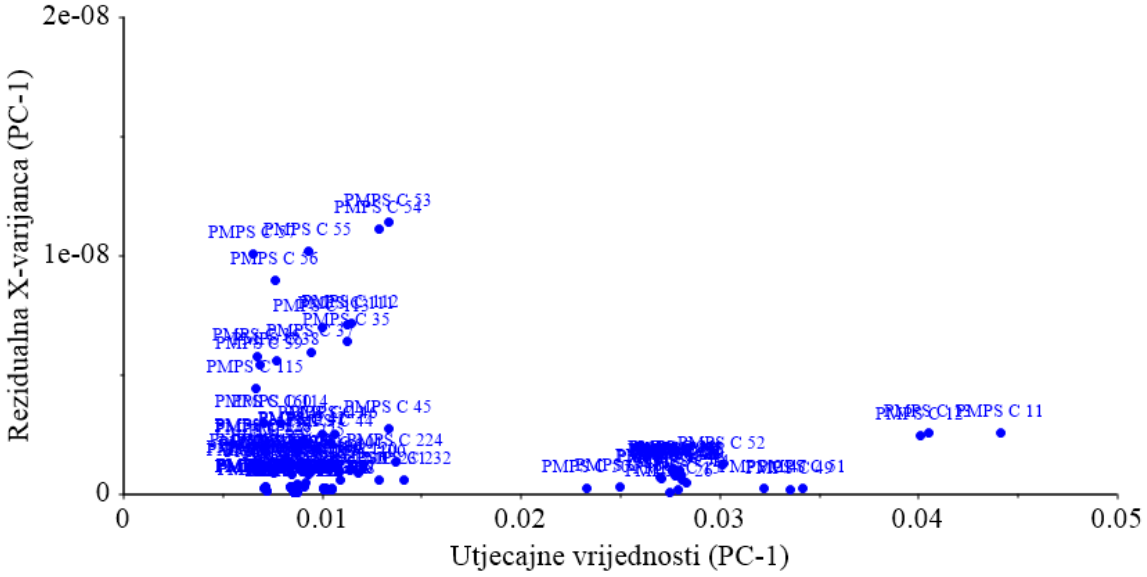
Opterećenja prikazana na Slikama 71.-74. ukazuju na najvažnije spektralne regije, koje definiraju PC1 i PC2. Spektralna područja odgovorna za definiranje prve glavne komponente (PC 1; Slika 72.) i druge glavne komponente (PC 2; Slika 73.) su: $\tilde{\nu}$ = 7000 - 6500 cm^{-1} proizlazi od prvog višeg tona O-H istezanja $\tilde{\nu}$ = 5970 - 5910 cm^{-1} koje proizlazi od prvog višeg tona acetamid metil C–H asimetričnog istezanja; $\tilde{\nu}$ = 5780 - 5840 cm^{-1} proizlazi od prvog višeg tona metilen C–H asimetričnog istezanja; spektralno područje $\tilde{\nu}$ = 5300 - 5100 cm^{-1} odgovara O-H kombinacijskim vibracijama; spektralno područje ν = 4900 - 4500 cm^{-1} proizlazi iz drugog višeg tona C=O istezanja, C-N istezanja i N-H savijanja u ravnini spektralno područje $\tilde{\nu}$ = 4360 - 4450 cm^{-1} odgovara kombinaciji C–H istezanja i CH_2 deformacije regija $\tilde{\nu}$ = 4060 - 4150 cm^{-1} odgovara kombinaciji istezanja C- H, C-C istezanja i C-O-C istezanja (Workman, 2001).

Kumulativnom kalibracijskom varijancom određen je optimalan broj PC-ova (Slika 75.), što je bio ključan preduvjet za formiranje kvalitetnog PCA modela.

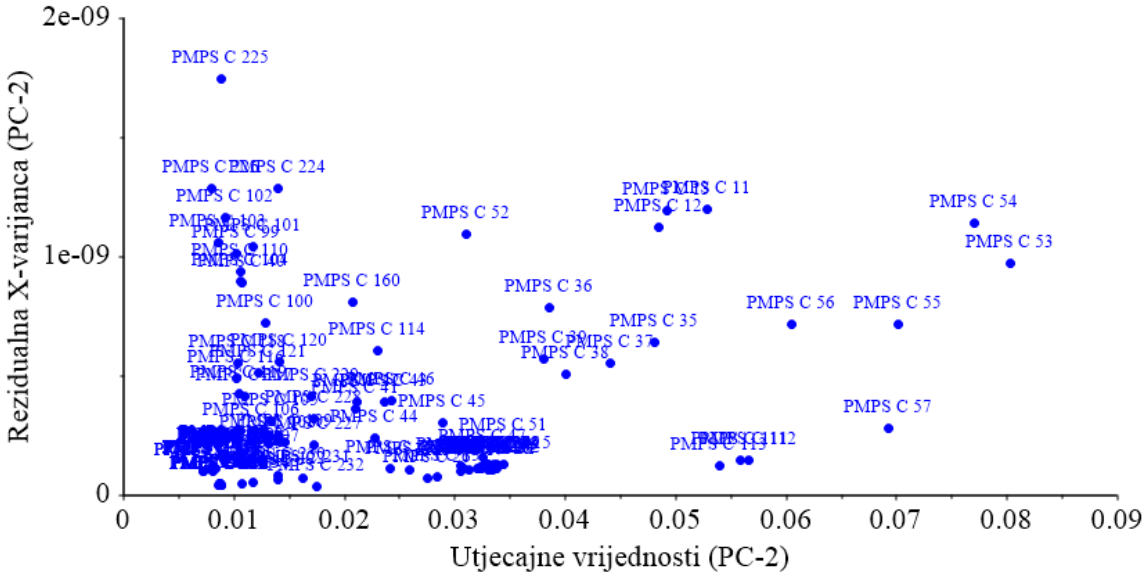


Slika 75. Kumulativna kalibracijska varijanca za svaki PC.

Slika 75. prikazuje koliko varijance opisuju različite glavne komponente (PC). Dvije glavne komponente (PC1 i PC2) opisuju ukupno 97 % ukupne varijance (PC1 - 84 % i PC2 - 13 %). Prisutnost netipičnih uzoraka u modelu je procijenjena analizom NIR spektralnih podataka pomoću utjecajnih vrijednosti, Hotelling T^2 statistike i Q-reziduala (ovdje ispod). (Slike 76. - 81.).



Slika 76. Rezidualna X-varianca i utjecajne vrijednosti uzoraka PMPS C za PC1.



Slika 77. Rezidualna X-varianca i utjecajne vrijednosti uzoraka PMPS C za PC2.

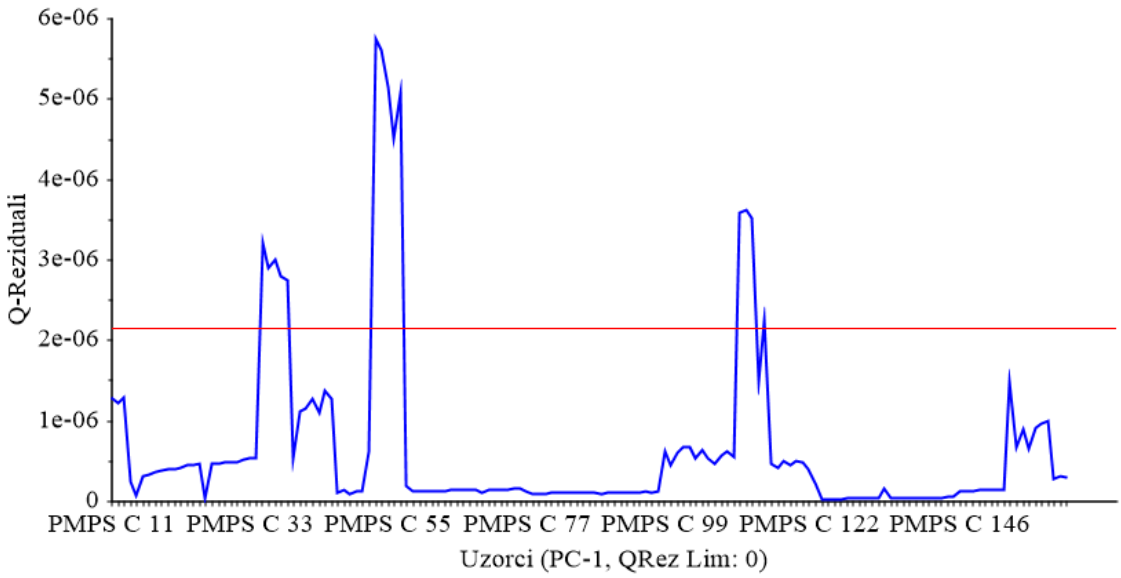
Na Slikama 76. i 77., redom, su prikazane rezidualne X-varijance i utjecajne vrijednosti za PC1 i PC2. Ove slike prikazuju koliko preostala X-varijanca pojedinog uzorka varira u usporedbi s drugim uzorcima i koliko je uzorak udaljen od modela. Kao što se može vidjeti na Slikama 76. i 77., osim uzoraka PMPS C 53 i PMPS C 54, većina uzoraka se relativno dobro uklapa u kategoriju PMPS C. Uzorci PMPS C 53 i PMPS C 54, imaju i visoku rezidualnu X-varijancu za PC1, i veliku rezidualnu varijancu i visoki utjecaj na model za PC2, što ukazuje na to da

predstavljaju izrazito opasne uzorke u kalibracijskom setu jer osim što ih model jako slabo opisuje imaju istovremeno visok utjecaj na model. Varijanca PC2 velikim dijelom potječe od upravo ovih uzoraka što posljedično tome dovodi vjerojatno do lose prediktivne sposobnosti budućeg modela. Radi svega navedenog potrebno je dodatno statističi analizirati uzorke PMPS C53 i PMPS C 54 kao potencijalne jako opasne netipične uzorke.

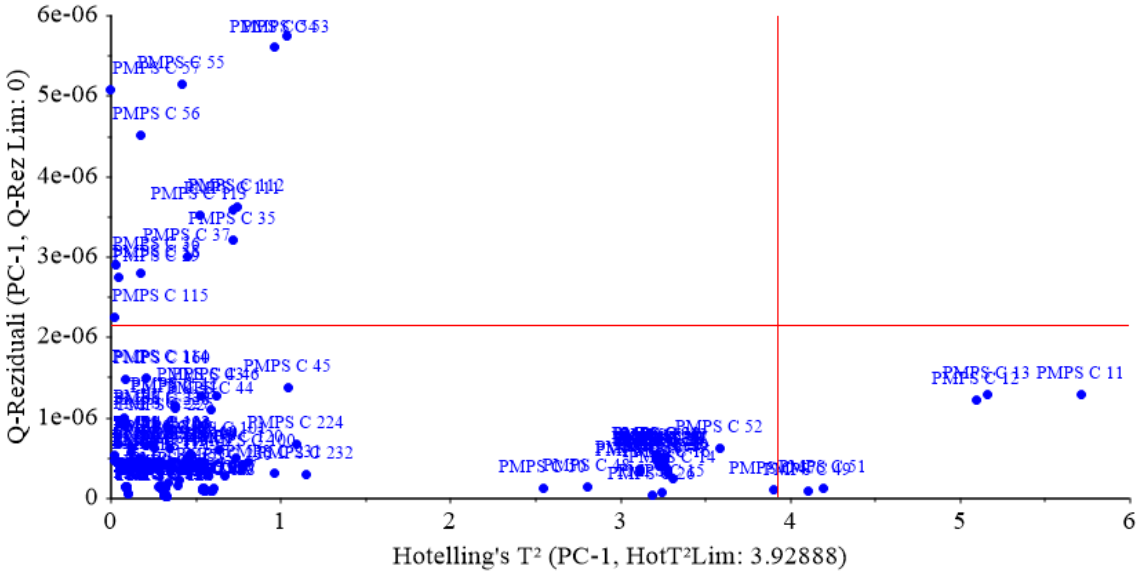
Analiza uzoraka visokih rezidualnih vrijednosti je važna za identificiranje uzoraka koji su udaljeni od središta unutar prostora koji opisuje PCA model (Slike 76. i 77.). Uzorci sa visokom rezidualnom vrijednosti te uzorci sa visokom utjecajnom vrijednosti koji se razlikuju od prosječnih uzoraka su potencijalni netipični uzorci, koje bi trebalo izuzeti iz formiranja PCA modela. Uzorci koje model dobro opisuje a imaju jako veliki utjecaj na model treba dobro ispitati upravo zbog svog velikog utjecaja na model.

Kada se procjenjuju netipične vrijednosti, dodatno se analiziraju vrijednosti Hotelling T^2 i Q-reziduali svakog pojedinog uzorka unutar skupa uzoraka (slike ovdje ispod).

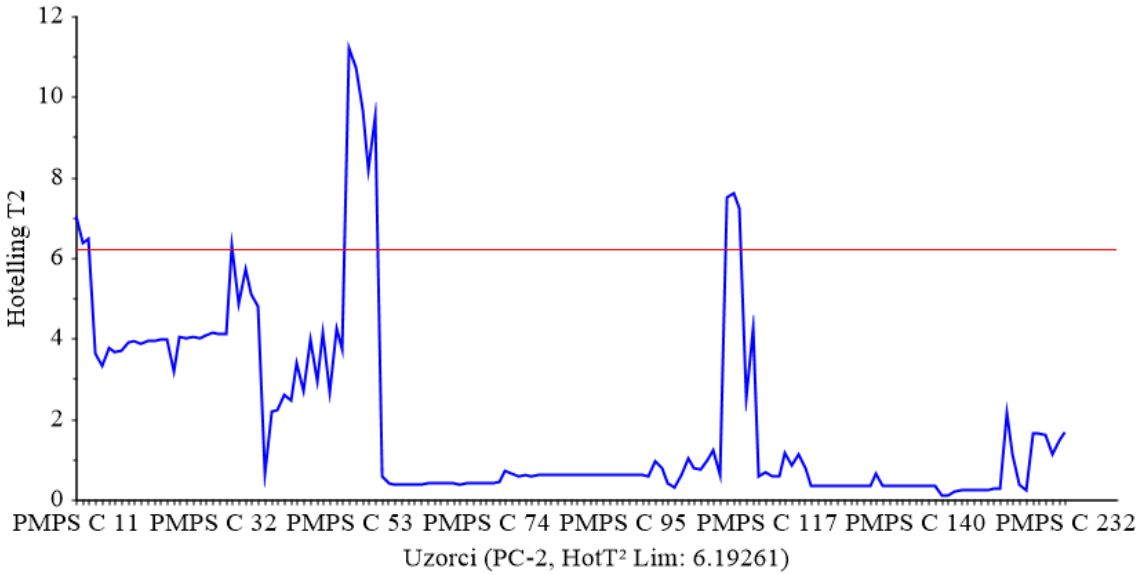
Identifikacija nepoznatih uzoraka PMPS A i PMPS C SIMCA metodom može se postići jedino kad uzorci iz kalibracijskog (trening) seta, koji se koriste u formiranju SIMCA modela, ne sadrže netipične uzorke. Pomoću Hotelling T^2 statistike i Q-reziduala se statistički mogu dodatno identificirati netipični uzorci unutar kalibracijskog (trening) seta uzoraka PMPS C (Slike 78.-80.).



Slika 78. Q-reziduali uzoraka PMPS C za PC1 s pripadajućom graničnom linijom (crvena linija).



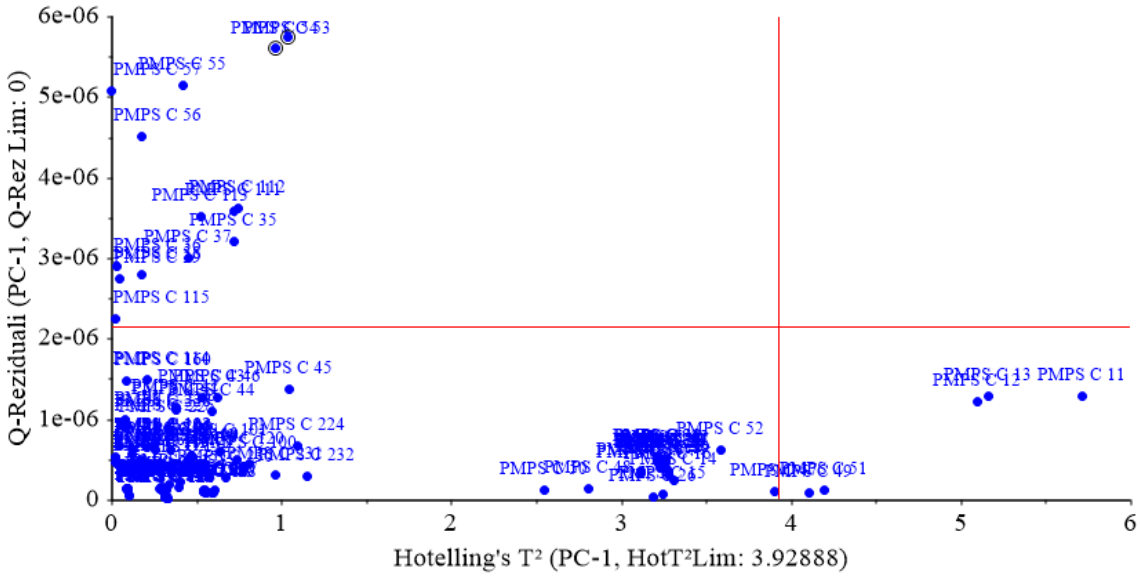
Slika 79. Hotelling T^2 statistika i Q-reziduali uzoraka PMPS C za PC1 s pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).



Slika 80. Hotelling T^2 statistika uzoraka PMPS C za PC2 sa pripadajućom kritičnom vrijednosti (crvena linija).

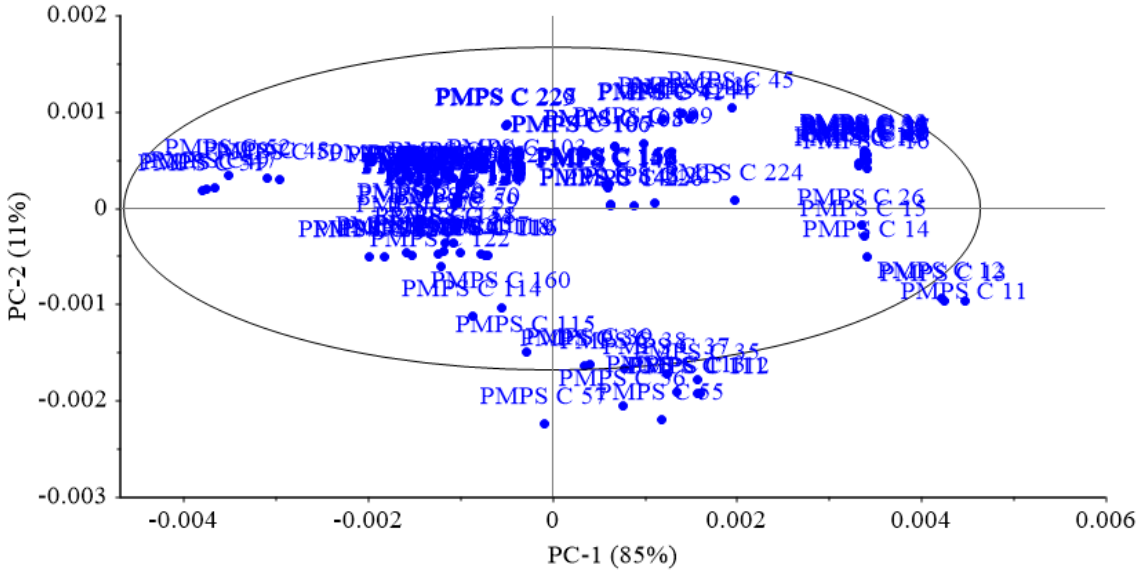
Na slikama 78. - 80. može se vidjeti da uzorci PMPS C 53 i PMPS C 54 imaju više vrijednosti Q reziduala i Hotelling T^2 u odnosu na ostale uzorke, te više od granične vrijednosti izračunate za Q reziduale i Hotelling T^2. Statističkom analizom eksperimentalno dobivenih NIR spektralnih podataka za PMPS C, utvrđeno je da svakako treba izdvojiti uzorke PMPS C 53 i

PMPS C 54 kao netipične uzorke iz kalibracijskog skupa uzoraka PMPS C. Vizualnim pregledom preostalih ekstremnih uzoraka primijećene su različite veličine čestica, te različita gustoća uzoraka u odnosu na ostale uzorke unutar kalibracijskog seta te su preostali uzorci identificirani kao ekstremni uzorci poželjni za formiranje modela te se neće izdvojiti kao netipični uzorci već će se zadržati u kalibracijskom skupu uzoraka. Također je ustanovljeno i da su unutar kalibracijskog seta, identificirani ekstremni uzorci PMPS C 55, PMPS 56 i PMPSC 57 sa visokim Q rezidualima. Ovo su uzorci dugotrajne stabilitetne studije koje je svakako poželjno i potrebno uključiti unutar kalibracijskog skupa uzoraka te ih se neće izuzeti iz kalibracijskog skupa. Zbog toga je bilo potrebno nanovo načiniti PCA model preostalih NIR spektralnih podataka PMPS C, ali ovaj put bez netipičnih uzoraka. U ponovljenoj statističkoj obradi označeni su netipični uzorci (Slika 81.). Ova se ponovljena statistička obrada može usporediti sa podacima na Slici 79.



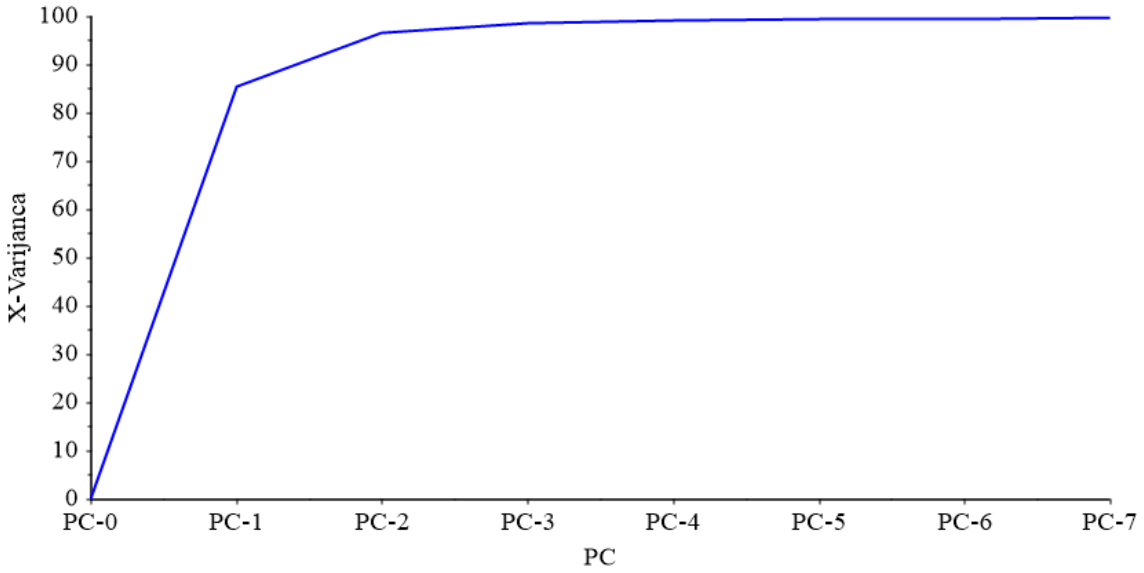
Slika 81. Hotelling T^2 statistika i Q-reziduali uzoraka PMPS C za PC1 s označenim netipičnim uzorcima i pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).

Nakon isključivanja netipičnih uzoraka ponovljena je PCA analiza NIR spektralnih podataka PMPS C (slika dolje).



Slika 82. Raspodjela faktorskih bodova PC1 i PC2 matematički obrađenih (Savitzky-Golay glačanje 3.9 s drugom derivacijom) NIR spektara PMPS C u području $\tilde{\nu}$ = 7768 - 3695 cm^{-1}. Elipsa označava 95 %-tni interval pouzanosti (Hotelling T^2 statistika).

Kumulativnom kalibracijskom varijancom određen je optimalan broj PC-ova (Slika 83.), što je preduvjet za formiranje kvalitetnog PCA modela.



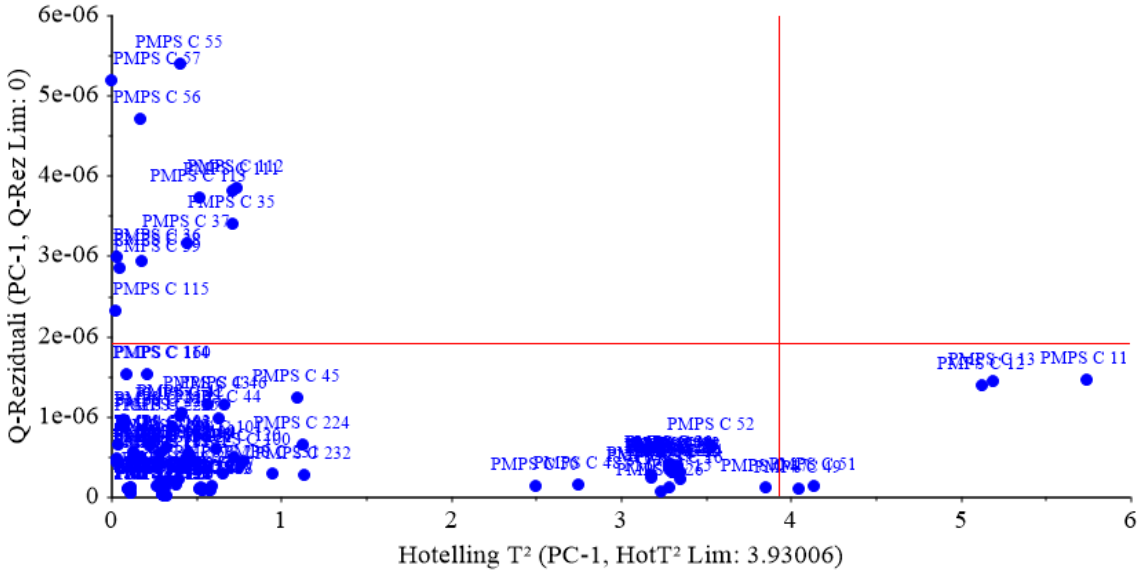
Slika 83. Kumulativna kalibracijska varijanca za svaki PC.

Tablica 2. Kumulativna kalibracijska varijanca za svaki PC nakon uklanjanja netipičnih uzoraka.

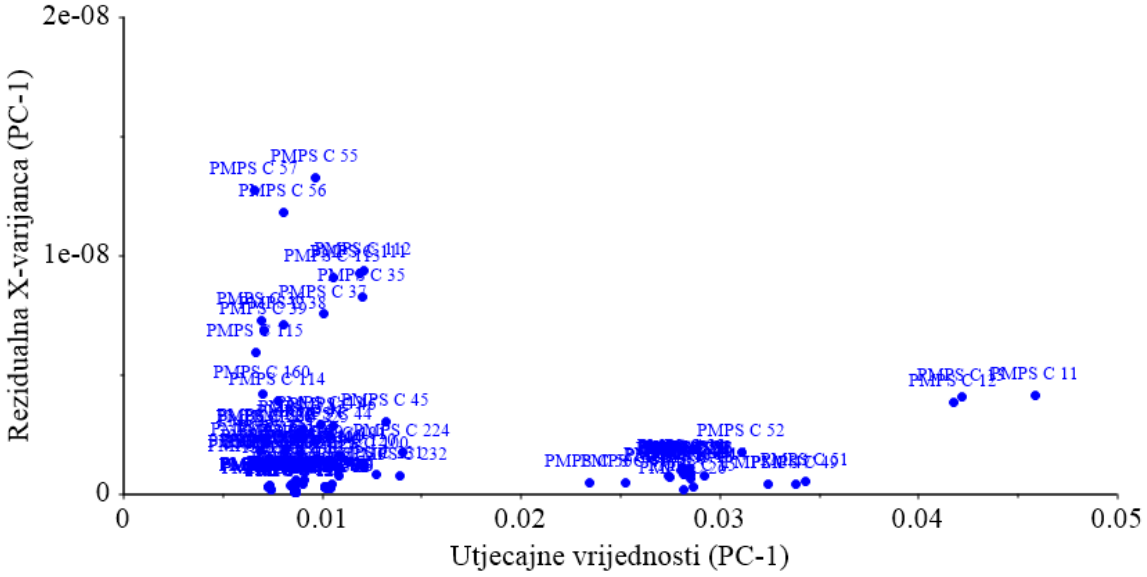
	PC 0	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6
Kalibracija	0	85.3972	95.8018	97.8457	98.6160	98.9457	99.1953

Iz Tablice 2. se može vidjeti da dva PC-a obuhvaćaju više od 95 % ukupne varijance, dok 3 PC-a ne obuhvaćaju značajno više, tj. obuhvaćaju nešto manje od 98%, te su kao optimalan broj za optimalnu dimenzionalnost modela odabrana dva PC-a

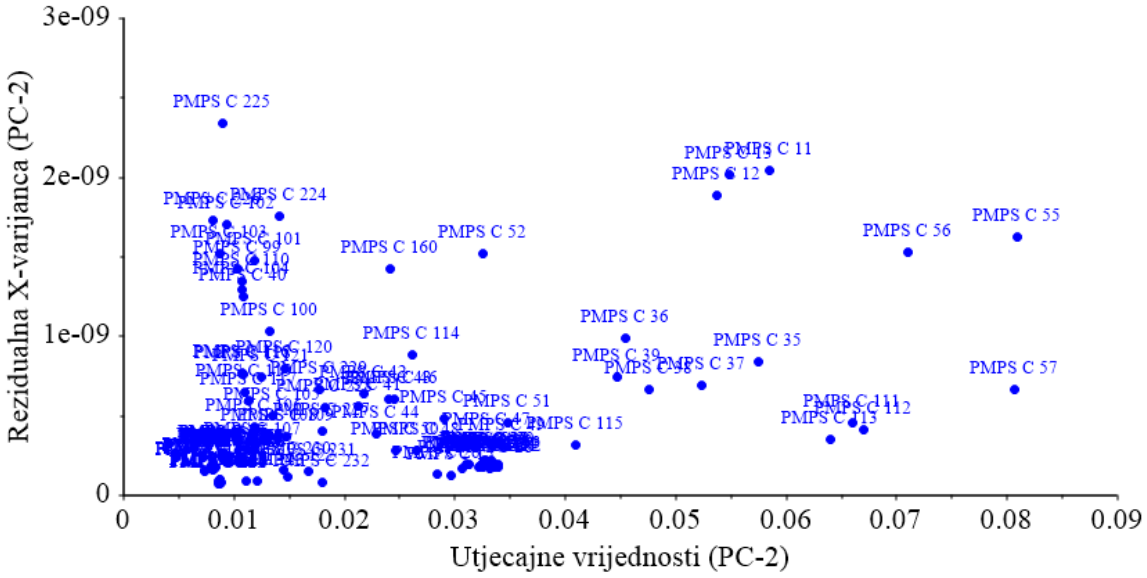
Također, načinio se: prikaz utjecajnih vrijednosti, Hotelling T^2 statistika i Q-reziduali sve u cilju identificiranja netipičnih PMPS C uzoraka.



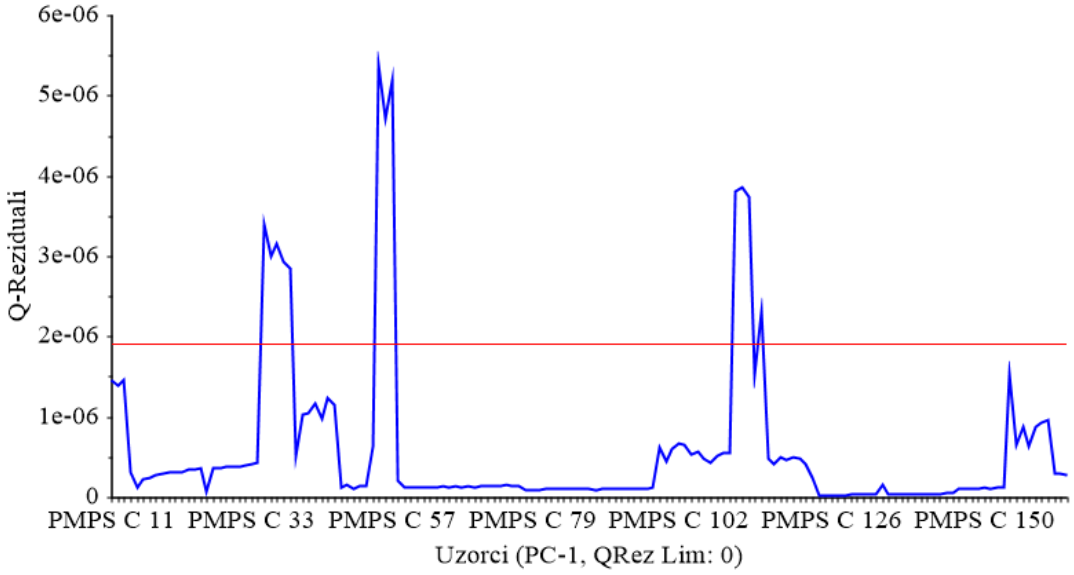
Slika 84. Hotelling T^2 statistika i Q-reziduali uzoraka PMPS C za PC1 sa pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).



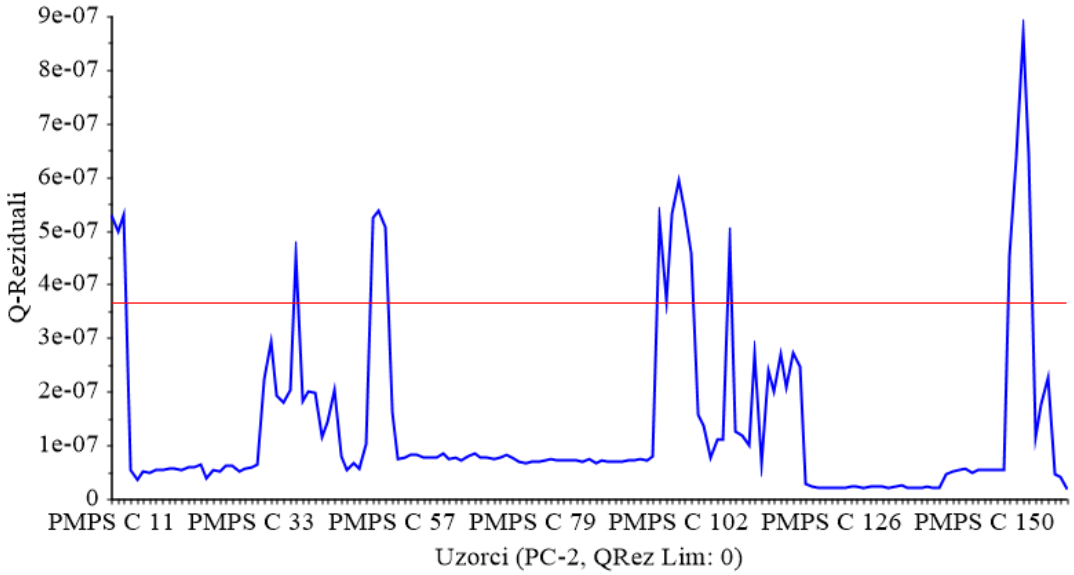
Slika 85. Rezidualna X-varijanca i utjecajna vrijednost uzoraka PMPS C za PC1.



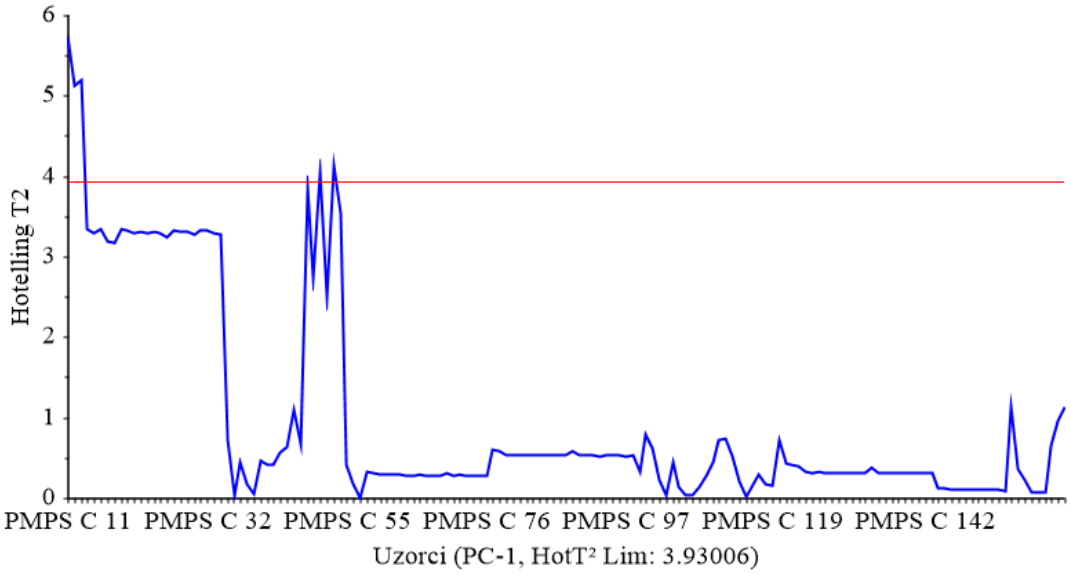
Slika 86. Rezidualna X-varijanca i utjecajna vrijednost uzoraka PMPS C za PC2.



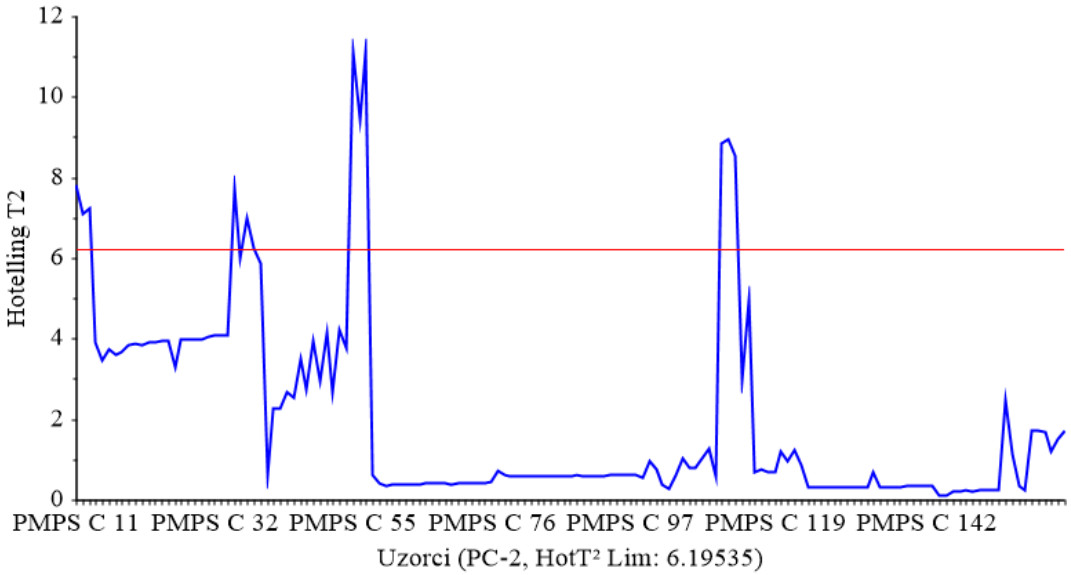
Slika 87. Q reziduali uzoraka PMPS C za PC1 s pripadajućom graničnom vrijednosti (crvena linija).



Slika 88. Q reziduali uzoraka PMPS C za PC2 s pripadajućom graničnom vrijednosti (crvena linija).



Slika 89. Hotelling T^2 statistika uzoraka PMPS C za PC1 sa pripadajućom kritičnom vrijednosti (crvena linija).



Slika 90. Hotelling T^2 statistika uzoraka PMPS C za PC2 sa pripadajućom kritičnom vrijednosti (crvena linija).

Prikazane utjecajne vrijednosti, Hotelling T^2 statistika i Q-reziduali (Slike 84. - 90.) prikazuju koliko su dobro uzorci kalibracijskog seta uzoraka PMPS C u skladu s formiranim PCA modelom tj. varijaciju izvan modela, udaljenost uzoraka do centra modela, odnosno varijaciju unutar modela te utjecaj uzoraka na PMPS C model, nakon izuzimanja netipičnih uzoraka.

Može se vidjeti prisutnost uzoraka viših Hotelling T^2 vrijednosti, i uzoraka viših vrijednosti Q reziduala. Ovi ekstremni uzorci od ključne su nam važnosti za formiranje robustnog i visoko učinkovitog modela. Vizualnim pregledom uzoraka, te pregledom sirovih NIR spektralnih podataka, utvrđeno je da su ovi uzorci realne proizvodne serije s uključenim proizvodnim varijabilnostima koje proizlaze iz složenosti biotehnološkog procesa proizvodnje, te uzorci stabilitetnih studija i poželjno je da su uključeni u formiranje modela, kako bi konačni model bio robustan i visokoučinkovit u identificiranju proizvodnih serija ovog meningokoknog polisaharida.

4.2.5.2. Optimizacija i validacija NIR SIMCA modela (Savitzky - Golay glačanje 3.9 s drugom derivacijom)

Optimizacija i validacija SIMCA modela provedena je pomoću različitih setova uzoraka PMPS A, C, W135 i Y, kako je opisano u poglavljima 4.2.5.2.1. i 4.2.5.2.2.

4.2.5.2.1. Optimizacija NIR SIMCA modela

Optimizacija ovog NIR SIMCA modela provedena je pomoću uzoraka iz test seta 1 i to na PCA modelima za PMPS A i C, koju su formirani s pomoću uzoraka iz trening seta. Kako je optimalan broj PC-ova dva za PMPS A i za PMPS C, korištena su dva PC-a za optimiranje NIR SIMCA modela i to za svaku serogrupu polisaharida. U test setu 1 su osim polisaharida serogrupe A i C još NIBCS standardi PMPS A i C i negativne probe polisaharida - PMPS W135 i Y. Uporabom NIBSC standarda provjereno je dali je uspješno ekstrahirana kemijska struktura proizvodnih uzoraka PMPS A i C, odnosno, da li će formirani SIMCA model ispravno identificirati oba polisaharida - PMPS A i C. Također je provjereno hoće li formirani NIR SIMCA model identificirati negativne probe - PMPS W135 i Y kao jedan (ili oba) od dva polisaharida serogrupa A i C. Rezultati dobiveni za NIR SIMCA model sa dva PC-a za svaku PMPS prikazani su u matrici zabune (Tablica 3) Izbor negativnih proba (PMPS W135 i Y) zasnovan je na konceptu najbližeg roda , gdje su negativne probe PMPS W135 i Y najbliži analozi PMPS A i PMPS C, proizvedeni na istom mjestu, istom tehnologijom te ispitani istim pravilima kontrole kvalitete.

Tablica 3. Matrica zabune validacijskih parametara za SIMCA model sa dva PC-a za svaku PMPS, koji su dobiveni za uzorke PMPS iz test seta 1.

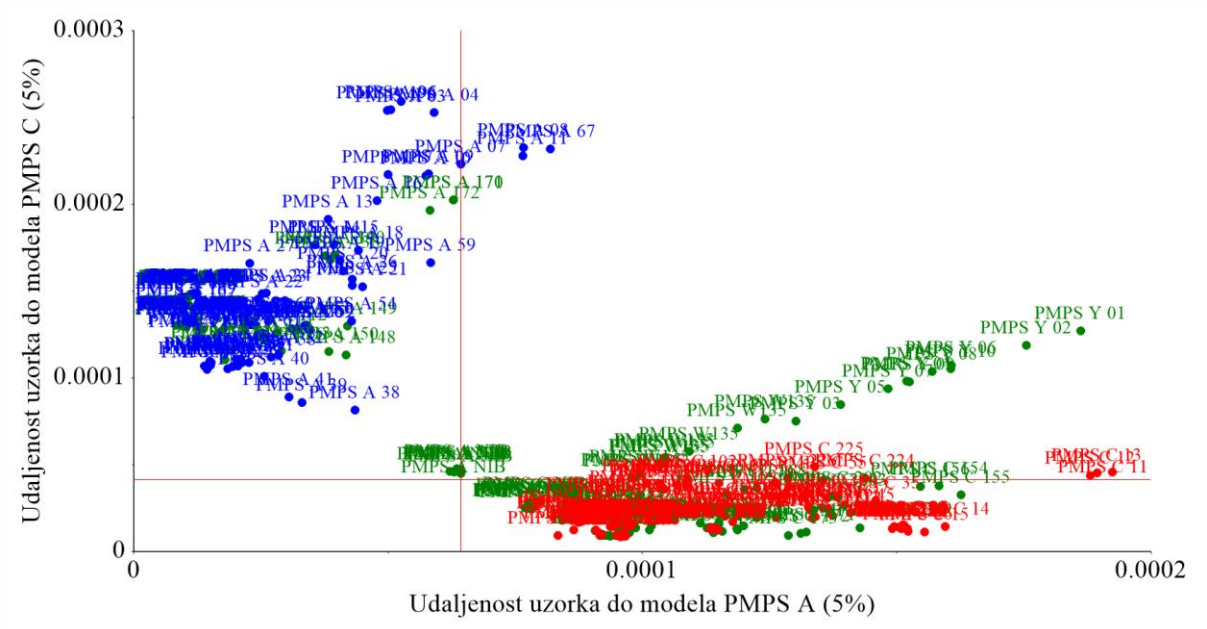
stvarno/predviđeno	klasa PMPS A	klasa PMPS C	nije klasificirano
klasa PMPS A	61	0	0
klasa PMPS C	0	98	0
klasa PMPS W135	0	0	10
klasa PMPS Y	0	0	10
CSNS	100%	100%	TSNS = 100%
CSPS	100%	100%	TSPS = 100%
CEFF	100%	100%	TEFF (za dva PC-a) = 100%

CSNS, CSPS, CEFF, TSNS, TSPS i TEFF su opisani u poglavlju 2.5.3.2.

Za procjenu izvedbenih sposobnosti NIR SIMCA identifikacijskog modela razmatrani su validacijski parametri, osjetljivost, specifičnost i učinkovitost za svaku ciljnu PMPS A i C kao i za ukupni SIMCA model. Rezultati dobiveni za negativne probe (PMPS W135 i PMPS Y) koriste se za demonstraciju specifičnosti formiranoga NIR SIMCA modela. Dobiveni rezultati pokazuju da SIMCA model sa dva PC-a sadrži sasvim dovoljno informacija koje obuhvaćaju heterogenost ciljnih PMPS i ispravno identificiraju sve uzorke u ciljne grupe, što znači da je ukupna osjetljivost SIMCA modela 100 %. Ovaj NIR SIMCA model zadržava uzorke negativnih proba neidentificiranim, pa je ukupna specifičnost SIMCA modela 100 %. Također formirani NIR SIMCA model pokazuje dobru separaciju ovih PMPS. Maksimalna učinkovitost NIR SIMCA modela iznosi 100 %. Reproducibilnost NIR SIMCA modela je testirana s pomoću vanjskog validacijskog seta uzoraka PMPS i to kako bi se potvrdila predikcijska sposobnost ovog modela (Poglavlje 4.2.5.2.2.).

Cooman dijagramom za SIMCA identifikacijski model, koji se temelji na prethodno definiranim PCA modelima za PMPS A i PMPS C pri 5 %-tnom nivou pouzdanosti, prikazana je udaljenost svakog uzorka od oba modela (PMPS A i PMPS C; Slika 91.). Tako je vizualizirano koliko su međusobno ova dva modela (PMPS A i PMPS C) različiti jedan od drugog. Novo klasificirani uzorci iz test seta 1 prikazani su na Slici 91. zelenom bojom, dok su uzorci iz kalibracijskog seta za dva modela (PMPS A i PMPS C) prikazani plavom (PMPS A) i crvenom bojom (PMPS C). Uzorak koji se klasificira kao PMPS A nalazi se s lijeve strane

okomite crte (plavo). Uzorak koji se klasificira kao PMPS C nalazi se ispod ispod vodoravne crte (crveno), koja predstavlja granicu za identifikaciju ovog PMPS. Uzorci koji bi se nalazili istovremeno lijevo od okomite crte i ispod vodoravne crte mogli bi biti članovi bilo koje skupine - PMPS A ili PMPS C, dok se uzorci koji bi se nalazili istovremeno desno od okomite crte i iznad vodoravne crte ne mogu identificirati kao PMPS A ili C.

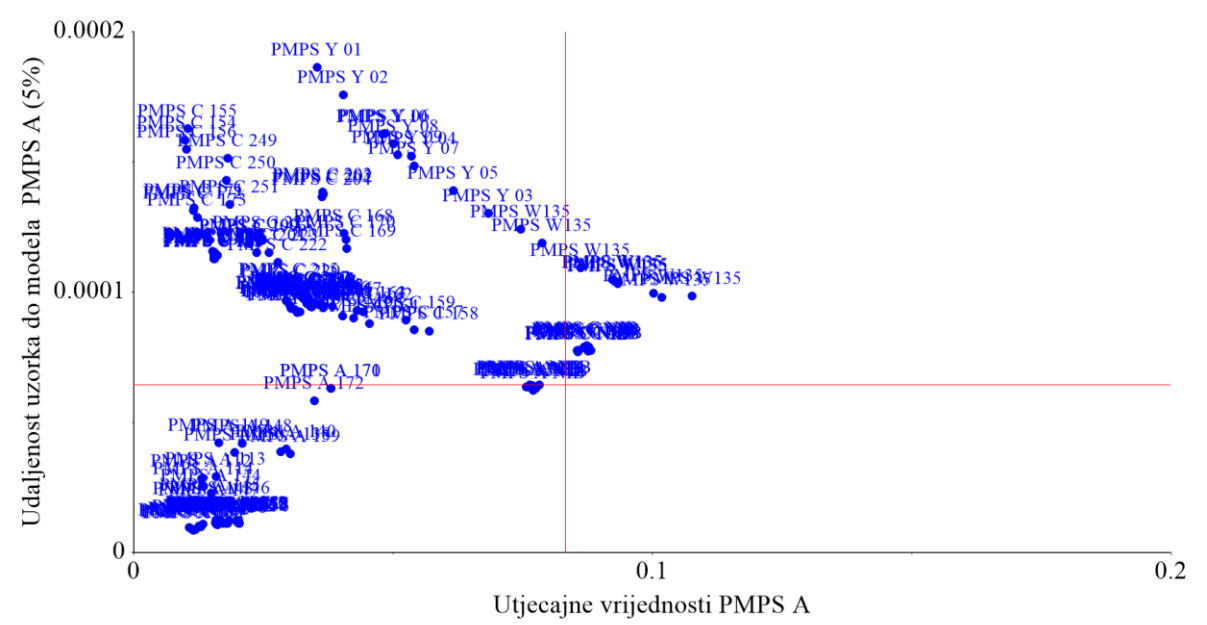


Slika 91. Cooman-ov dijagram s pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).

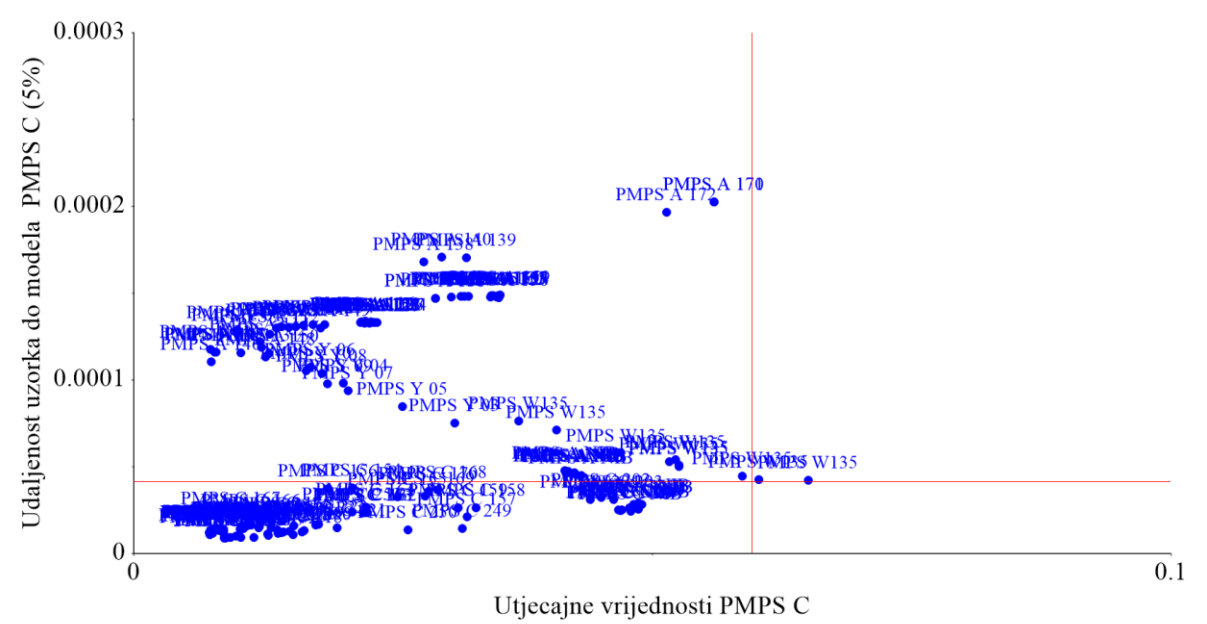
Dakle, formirani NIR SIMCA model je uspješno identificirao sve uzorke iz test seta 1, a posebno su pomoću ovoga SIMCA modela ispravno identificirani NIBSC standardi oba polisaharida - PMPS A i C. Osim toga, PMPS W135 i Y (polisaharidi u svojoj kemijskoj strukturi slični polisaharidima serogrupe C) nisu identificirani pomoću ovog modela kao pripadnici PMPS A ili C. Uspješno je ekstrahirana kemijska struktura obaju polisaharida (PMPS A i C) iz NIR spektralnih podataka i to iz proizvodnih uzoraka ovih polisaharida. Na ovaj način potvrđena je hipoteza da model formiran na proizvodnim uzorcima uspješno identificira odgovarajuće NIBSC standarde. Iz Cooman dijagrama (Slika 91.) se jasno vidi da su svi uzorci PMPS iz ovog test seta 1 nedvosmisleno dodijeljeni pomoću NIR SIMCA modela u odgovarajuće serogrupe i da nema uzoraka koji su dvosmisleno identificirani u dvije serogrupe. Sve zajedno upućuje na visokoučinkovit NIR SIMCA model.

Svi postupci predobrade NIR spektara (druga derivacija, Savitzky-Golay glačanje), te odbacivanje netipičnih uzoraka, rezultirali su formiranjem NIR SIMCA modela s vrlo visokom

mogućnosti klasifikacije. Pored toga, udaljenosti između PCA modela svake serogrupe PMPS su relativno velike, što ukazuje na to da formirani NIR SIMCA model karakterizira izvrsna razlučivost među ovim polisaharidima. Nepoznati uzorci PMPS A i PMPS C se uspoređuju sa PCA modelom svake grupe polisaharida i to uz korištenje dvaju parametra - udaljenosti uzorka do modela (Si) i utjecajne vrijednosti uzoraka (Hi). Si je udaljenost uzorka do centra modela, a Hi opisuje koliko bi utjecaja imao uzorak na model kada bi bio uključen u model. Ova usporedba je načinjena za PMPS A (Slika 92.) i za PMPS C (Slika 93.).



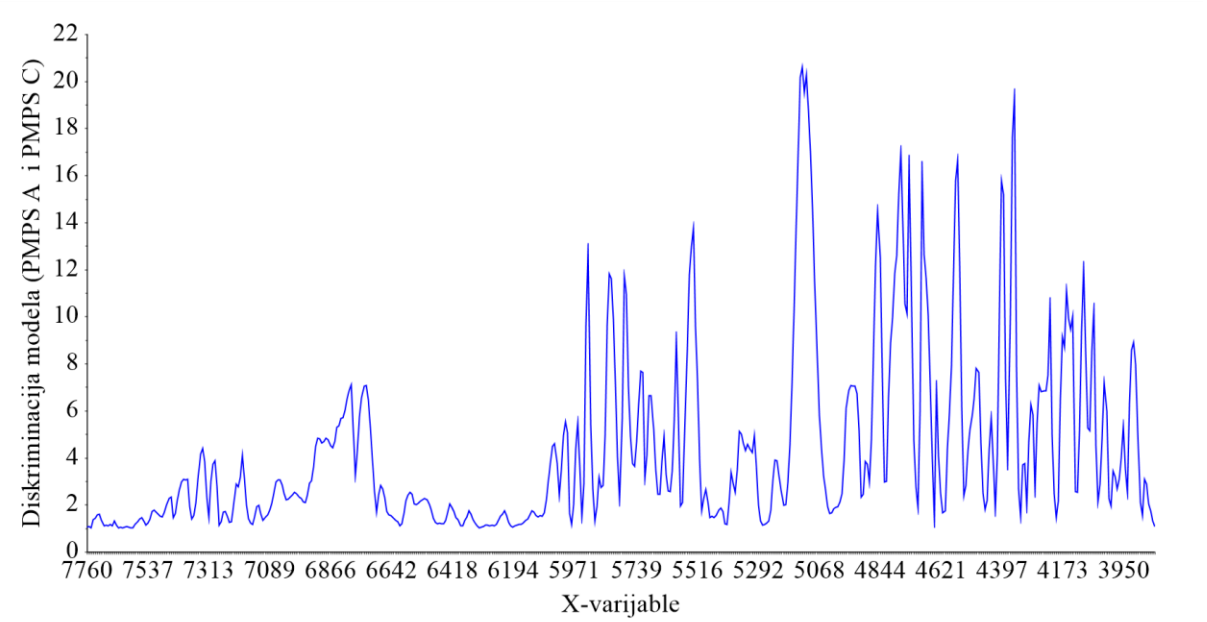
Slika 92. Udaljenosti uzorka PMPS A od modela (Si) s utjecajnim vrijednostima ovih uzoraka (Hi) za model PMPS A i pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).



Slika 93. Udaljenosti uzorka PMPS C od modela (Si) s utjecajnim vrijednostima ovih uzoraka (Hi) za model PMPS C i pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).

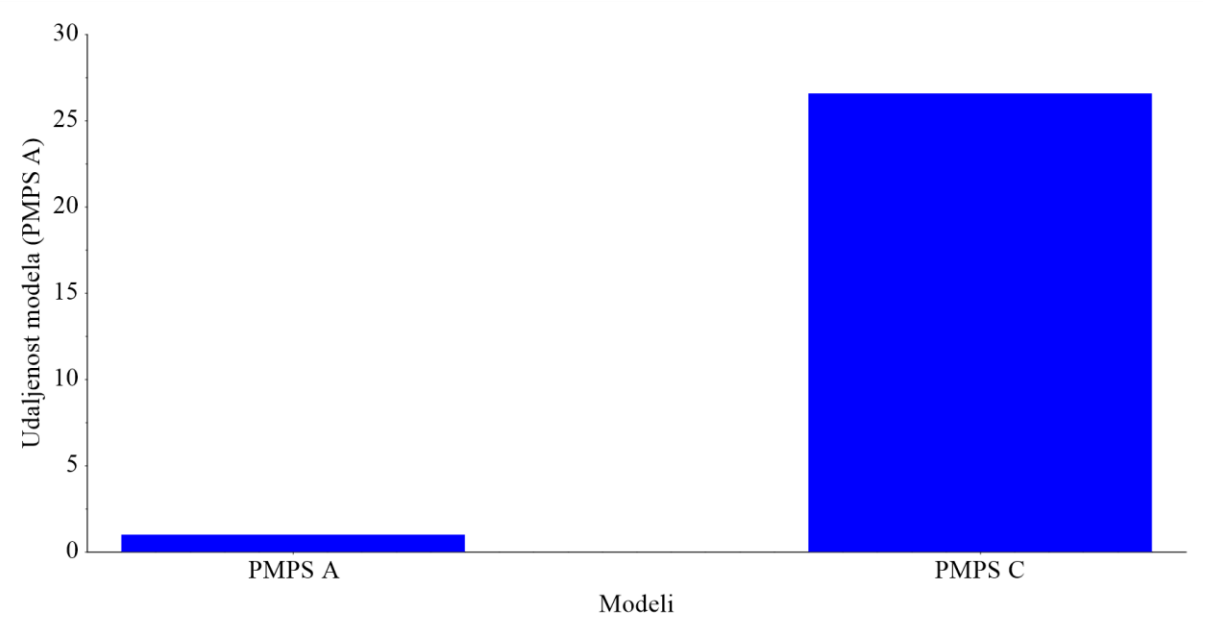
Slike 92. i 93. prikazuju udaljenost uzoraka PMPS A i C od PCA modela za obe serogrupe polisaharida u NIR SIMCA modelu. Uzorci jedne serogrupe iz kalibracijskog seta udaljeni su od PCA modela druge serogrupe i potpuno se razlikuju jedni od drugih. Nepoznati uzorci iz optimizacijskog test seta 1 su također ispravno dodjeljeni u odgovarajuću serogrupu pomoću formiranoga SIMCA modela. Izuzetno je važno da niti jedan uzorak PMPS nije identificiran i dodijeljen u dvije serogrupe istovremeno ili u pogrešnu serogrupu. Zbog toga, SIMCA identifikacija, koja se zasniva na PCA modeliranju se pokazala kao vrlo učinkovita u identifikaciji PMPS A i PMPS C.

Kako bi se definirale NIR spektralne regije, koje najviše doprinose diskriminaciji dviju serogrupa polisaharida, bilo je potrebno prikazati diskriminacijsku moć različitih valnih brojeva unutar snimljenih NIR spektara uzoraka iz kalibracijskog seta (Slika 94.).



Slika 94. Diskriminacijska moć različitih valnih brojeva ($\tilde{\nu}$) NIR spektara u diskriminaciji PMPS A i PMPS C.

Na slici 94. prikazana je diskriminacijska moć različitih valnih brojeva (ν) u diskriminaciji dviju serogrupa - PMPS A i PMPS C. Iako većina valnih brojeva ima utjecaj na razlikovanje uzoraka ovih PMPS, valni brojevi koji su imali najveći utjecaj na diskriminacju ovih serogrupa nalaze se u spektralnim regijama, kako slijedi: $\tilde{\nu}$ = 7000 - 6500 cm^{-1}; $\tilde{\nu}$ = 5970 - 5910 cm^{-1}; $\tilde{\nu}$ = 5780 - 5840 cm^{-1}; $\tilde{\nu}$ = 5150 - 5240 cm^{-1} ; $\tilde{\nu}$ = 4900 - 4500 cm^{-1} i $\tilde{\nu}$ = 4360 - 4450 cm^{-1} i $\tilde{\nu}$ = 4060 - 4150 cm^{-1}. Na temelju opisanih rezultata moglo bi se zaključiti da bi diskriminacija među serogrupama polisaharida mogla biti posljedica prvog višeg tona O-H vibracijskog istezanja; prvog višeg tona acetamid metil C–H asimetričnog istezanja; prvog višeg tona metilen C–H asimetričnog istezanja; kombinaciji polisaharidnog O-H istezanja, H-O-H deformacije, i O-H savijanja; drugog višeg tona C=O istezanja, C-N istezanja i N-H savijanja u ravnini; zatim kombinaciji C–H istezanja i CH_2 deformacije kao i kombinaciji C–H istezanja, C-C istezanja i C-O-C istezanja (Workman, 2001).



Slika 95. Udaljenost modela PMPS A od modela PMPS C

Tablica 4. Udaljenost dvaju PCA modela.

	PMPS A	PMPS C
PMPS A	1	26.57153
PMPS C	26.57153	1

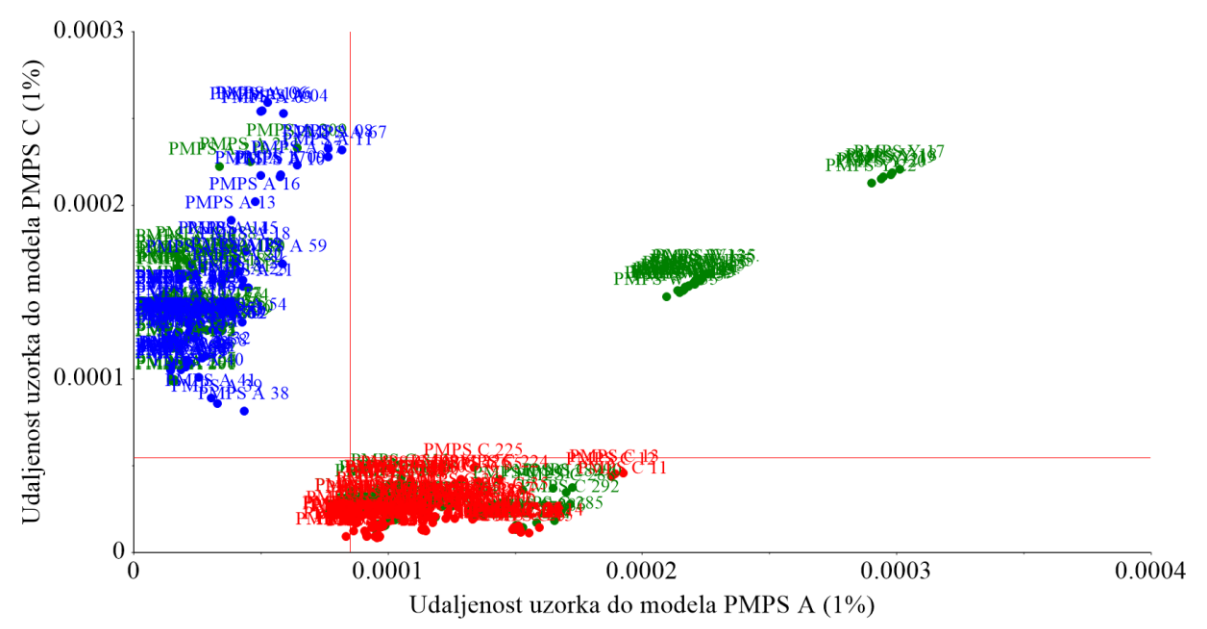
Slika 95. prikazuje udaljenost između dvaju PCA modela - PCA modela za PMPS A i PMPS C. Udaljenost modela kvantificira koliko su ova dva modela različita jedan od drugog. Veća udaljenost između dvaju modela ukazuje da su oba modela vrlo jasno odijeljena.

4.2.5.2.2. Validacija NIR SIMCA modela

4.2.5.2.2.1. Parametri validacije NIR SIMCA modela

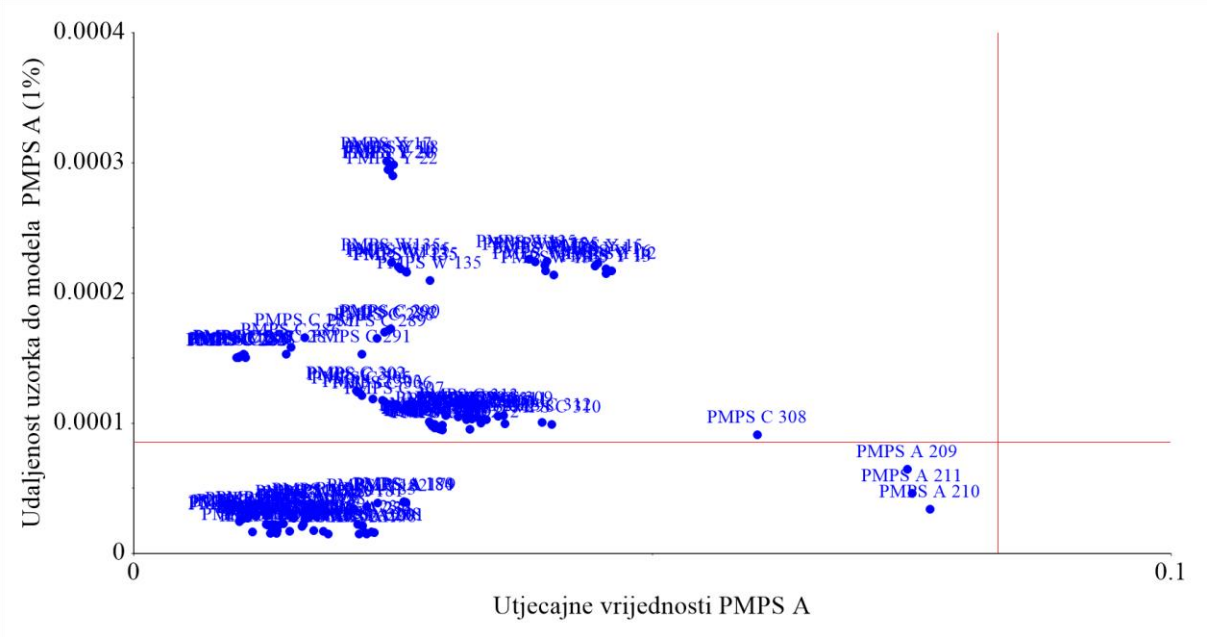
Nakon optimizacije formiranoga NIR SIMCA modela i to u cilju procjene identifikacijske sposobnosti ovoga modela s odabranim brojem glavnih komponenti za pojedini PCA model PMPS A i PMPS C, provedena je validacija formiranog NIR SIMCA modela i to pomoću vanjskog test seta uzoraka PMPS A, C, W135 i Y na PCA modelima PMPS A i PMPS C formiranim s pomoću uzoraka iz trening seta (poglavlje 4.2.5.2.2.).

Slika 96. prikazuje Cooman dijagram za identifikacijski NIR SIMCA model, koji se temelji na prethodno definiranim PCA modelima za PMPS A i PMPS C, pri čemu je granica pouzdanosti 1 %.

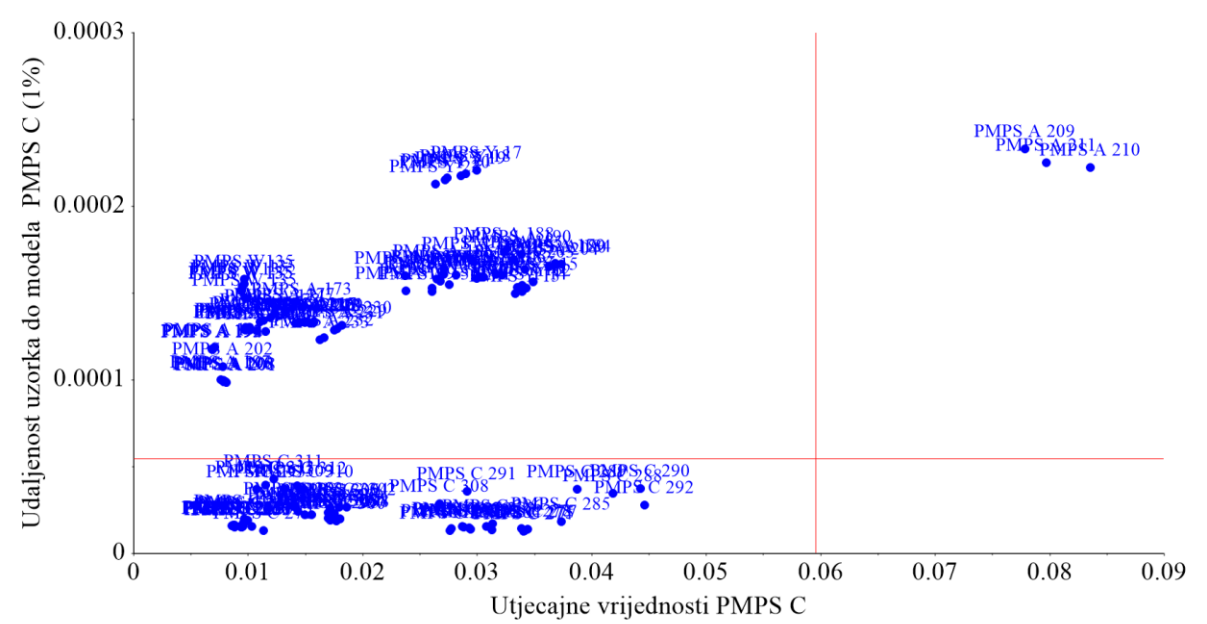


Slika 96. Cooman dijagram s pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).

Na Cooman dijagramu su uključene granice pripadnosti za oba modela, pa se na taj način može vidjeti pripada li uzorak u jednu klasu, obje klase ili ne pripada niti u jednu klasu PMPS. Iz ovog prikaza se jasno vidi da su svi uzorci nedvosmisleno dodijeljeni u odgovarajuće serogrupe (PMPS A, plavo; ili PMPS C, crveno) i da nema uzoraka između dvije serogrupe, što potvrđuje učinkovitost NIR SIMCA modela. Uzorci PMPS W135 i PMPS Y (zeleno), koji su korišteni kao negativne probe, očekivano nisu dodijeljeni niti u jednu klasu PMPS.



Slika 97. Udaljenosti uzorka PMPS A od modela (Si) s utjecajnim vrijednostima ovih uzoraka (Hi) za model PMPS A i pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).



Slika 98. Udaljenosti uzorka PMPS C od modela (Si) s utjecajnim vrijednostima ovih uzoraka (Hi) za model PMPS C i pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).

Prikaz udaljenosti uzoraka od modela (Si) s utjecajnim vrijednostima (Hi) za modele PMPS A i C sa postavljenim granicama za ove obje vrijednosti daje izvrstan uvid u ispravnu klasifikaciju PMPS u odgovarajući model. Na slikama 96-98 može se vidjeti da su svi (nepoznati) uzorci iz vanjskog validacijskog seta ispravno dodjeljeni u odgovarajuću serogrupu. Niti jedan uzorak nije kategoriziran u dvije serogrupe istovremeno niti je pogrešno identificiran kao pripadnik druge serogrupe, kao što je dodatno istaknuto u Tablici 5.

Tablica 5. Matrica zabune validacijskih parametara za SIMCA model sa dva PC-a za svaku serogrupu i ukupni NIR SIMCA model, kod kojih su korišteni uzorci iz vanjskog test seta.

stvarno/predviđeno	klasa PMPS A	klasa PMPS C	nije klasificirano
klasa PMPS A	60	0	0
klasa PMPS C	0	62	0
klasa PMPS W135	0	0	12
klasa PMPS Y	0	0	12
CSNS	100%	100%	TSNS = 100%
CSPS	100%	100%	TSPS = 100%
CEFF	100%	100%	TEFF (2 PC) = 100%

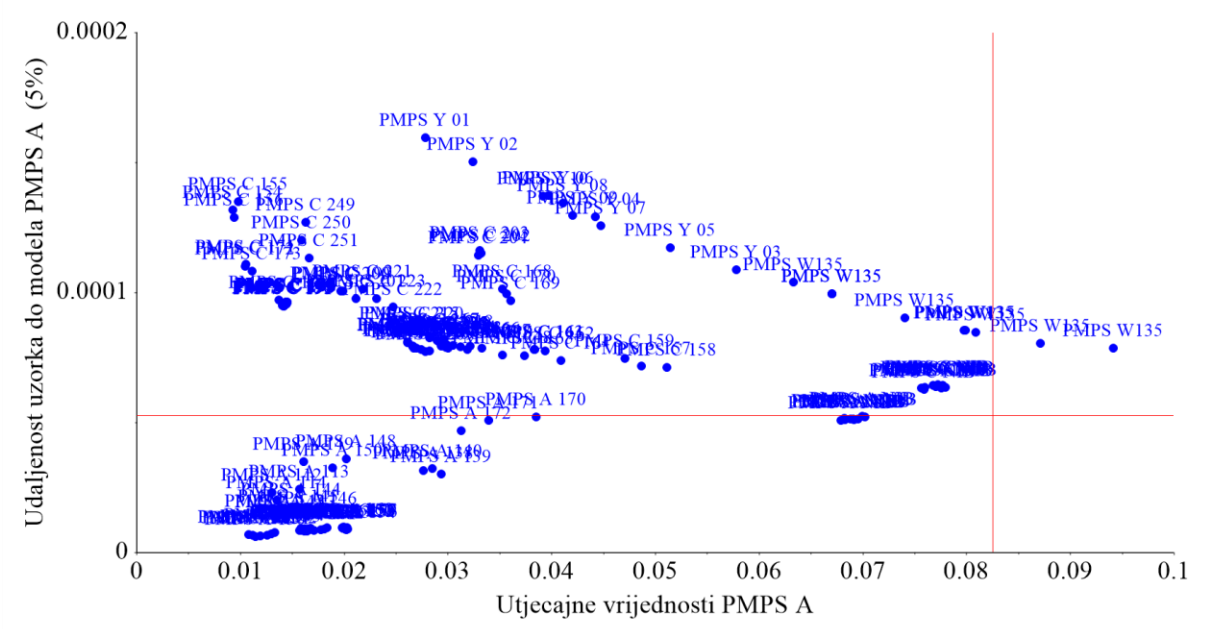
CSNS, CSPS, CEFF, TSNS, TSPS i TEFF su opisani u poglavlju 2.5.3.2.

Temeljem svih ovdje opisanih rezultata može se zaključiti kako je primjena NIR spektroskopije u kombinaciji sa formiranim SIMCA modelom, a koji se temelji na PCA modeliranju i to uz korištenje proizvodnih serija PMPS A, C, W135 i Y, visoko učinkovita za identifikaciju PMPS A i PMPS C.

4.2.5.3. Optimizacija i validacija NIR SIMCA jednoklasnog modela PMPS A (Savitzky - Golay glačanje 3.9 s drugom derivacijom)

Istražen je i pristup NIR SIMCA modela gdje je korišten samo formirani PCA model PMPS A, takozvani jednoklasni klasifikator, u autentifikacijske svrhe. Ovim pristupom identificiraju se meningokokni polisaharidi samo ciljne serogrupe dok se polisaharidi ostalih serogrupa ne identificiraju kao pripadnici klase (serogrupe).

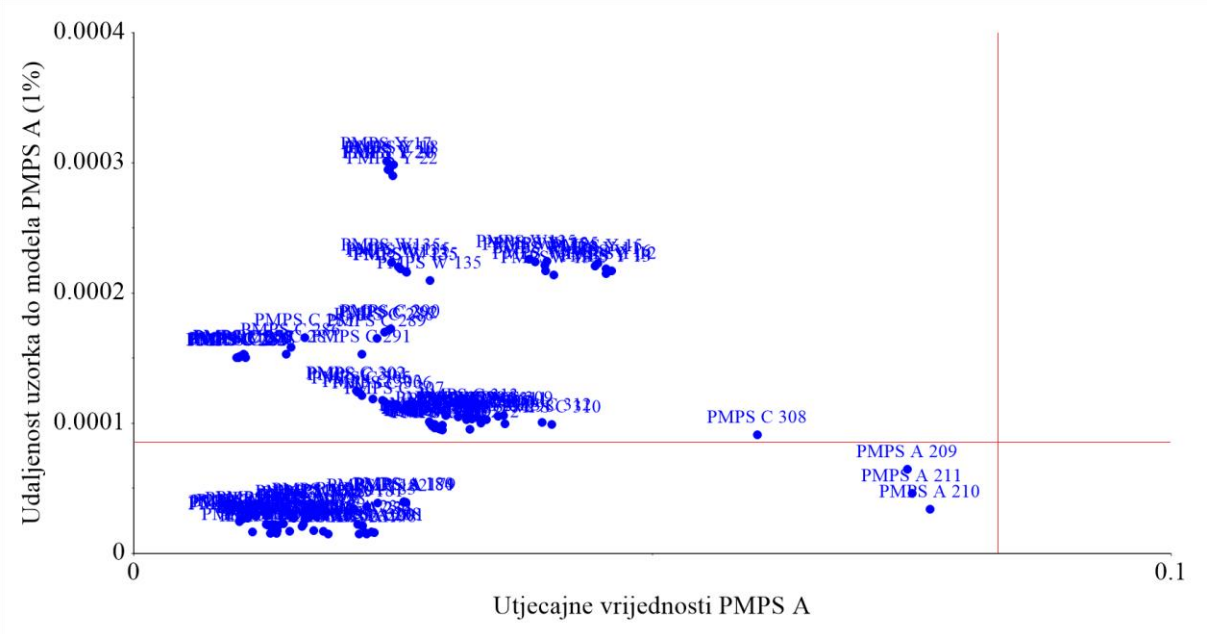
Optimizacija formiranog PMPS A NIR SIMCA modela u autentifikacijske svrhe provedena je test setom 1 (Slika 99.)



Slika 99. Udaljenosti uzorka PMPS A od modela (Si) s utjecajnim vrijednostima ovih uzoraka (Hi) za model PMPS A i pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).

Na slici 99. može se vidjeti da su svi (nepoznati) uzorci iz test seta 1 ispravno dodjeljeni u odgovarajuću serogrupu. Niti jedan uzorak PMPS C, PMPS W135 I PMPS Y, nije pogrešno identificiran kao pripadnik serogrupa PMPS A.

Validacija formiranog PMPS A NIR SIMCA modela u autentifikacijske svrhe provedena je vanjskim test setom (Slika 100.).

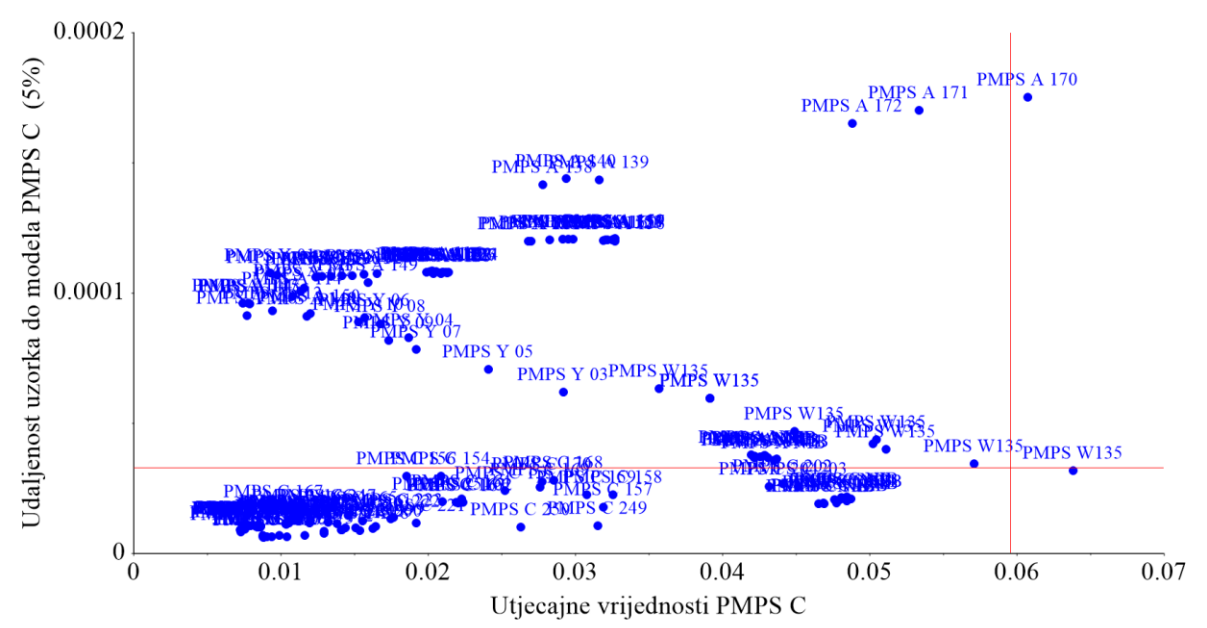


Slika 100. Udaljenosti uzorka PMPS A od modela (Si) s utjecajnim vrijednostima ovih uzoraka (Hi) za model PMPS A i pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).

Na slici 100. može se vidjeti da su svi (nepoznati) uzorci iz vanjskog test seta ispravno dodjeljeni u odgovarajuću serogrupu. Niti jedan uzorak PMPS C, PMPS W135 I PMPS Y, nije pogrešno identificiran kao pripadnik serogrupa PMPS A.

4.2.5.4. Optimizacija i validacija NIR SIMCA jednoklasnog modela PMPS C (Savitzky - Golay glačanje 3.9 s drugom derivacijom)

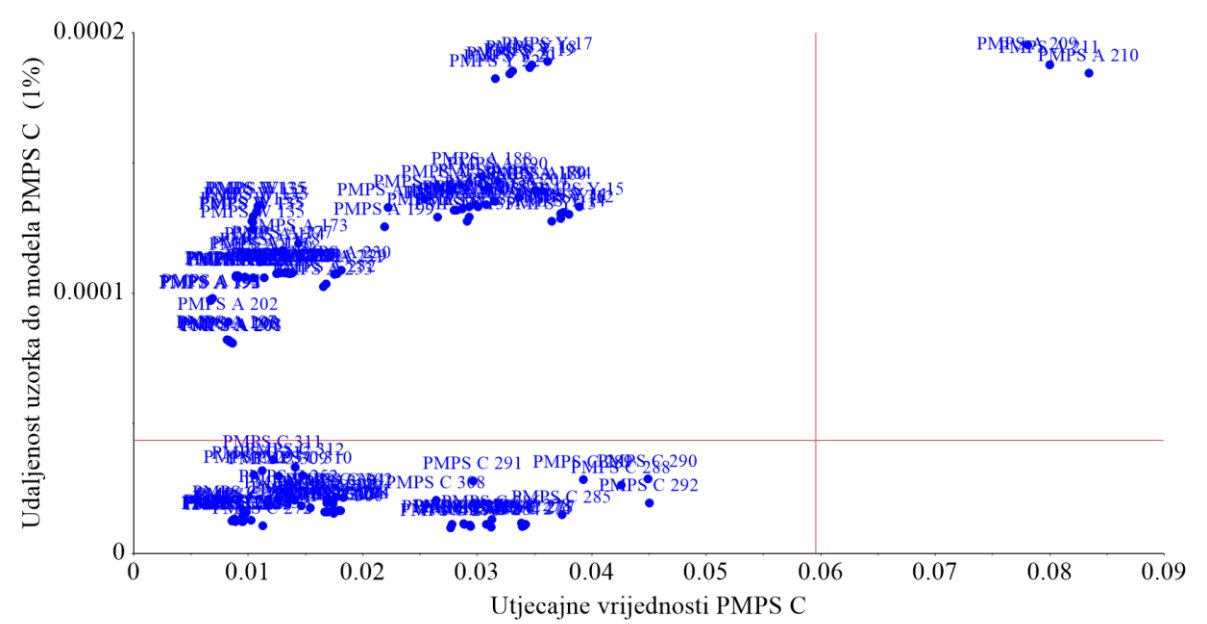
Istražen je i pristup NIR SIMCA modela gdje je korišten samo formirani PCA model PMPS C, takozvani jednoklasni klasifikator, u autentifikacijske svrhe.
Optimizacija formiranog PMPS C NIR SIMCA modela u autentifikacijske svrhe provedena je test setom 1 (Slika 101.).



Slika 101. Udaljenosti uzorka PMPS C od modela (Si) s utjecajnim vrijednostima ovih uzoraka (Hi) za model PMPS C i pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).

Na slici 101. može se vidjeti da su svi (nepoznati) uzorci iz test seta 1 ispravno dodjeljeni u odgovarajuću serogrupu. Niti jedan uzorak PMPS A, PMPS W135 I PMPS Y, nije pogrešno identificiran kao pripadnik serogrupa PMPS C.

Validacija formiranog PMPS A NIR SIMCA modela u autentifikacijske svrhe provedena je vanjskim test setom (Slika 102.).



Slika 102. Udaljenosti uzorka PMPS C od modela (Si) s utjecajnim vrijednostima ovih uzoraka (Hi) za model PMPS C i pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).

Na slici 102. može se vidjeti da su svi (nepoznati) uzorci iz vanjskog test seta ispravno dodjeljeni u odgovarajuću serogrupu. Niti jedan uzorak PMPS A, PMPS W135 I PMPS Y, nije pogrešno identificiran kao pripadnik serogrupa PMPS C.

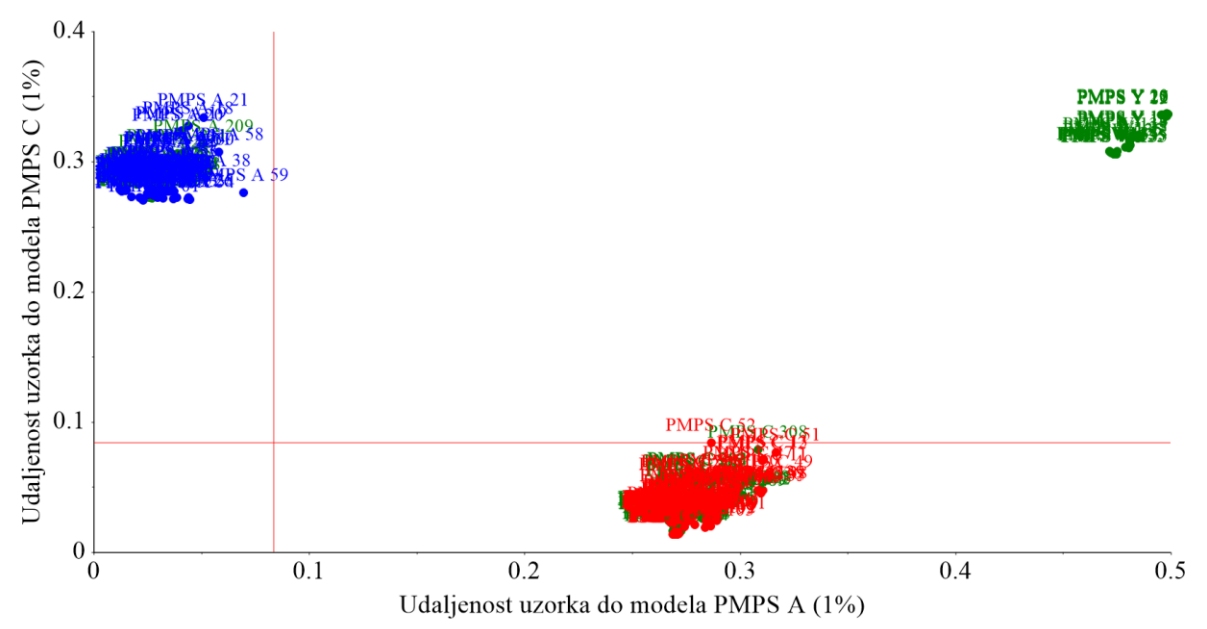
Iz dobivenih rezultata prkazanih na Slikama 99. - 102., može se zaključiti da je pristup NIR SIMCA modela kao jednoklasnog klasifikatora visoko učinkovit i robustan model u autentifikacijske svrhe ciljnog meningokoknog polisaharida.

4.2.5.5. Validacija NIR SIMCA modela (Savitzky - Golay glačanje 3.9 s drugom derivacijom i SNV)

Također je provedena i validacija SIMCA modela formiranog nakon matematičke obrade NIR spektara SNV-om i Savitzky-Golay glačanjem 3.9 s drugom derivacijom. Kako bi mogli usporediti rezultate validacije za NIR SIMCA modele dobivene obradom NIR spektralnih podataka različitom kombinacijom matematičkih predtretmana, prikazat će se rezultati validacije ovog formiranog NIR SIMCA modela. (ovdje ispod).

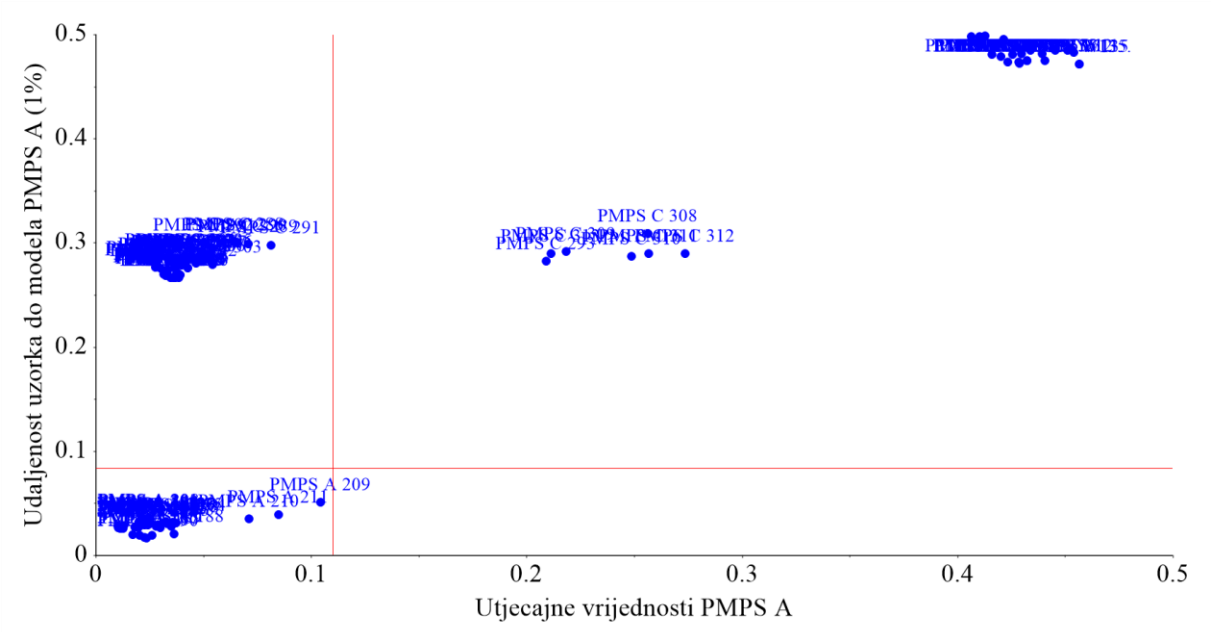
4.2.5.5.1. Parametri validacije NIR SIMCA modela

Slika 103. prikazuje Cooman dijagram za identifikacijski NIR SIMCA model, koji se temelji na prethodno definiranim PCA modelima za PMPS A i PMPS C, pri čemu je nivo značajnosti 1 %.

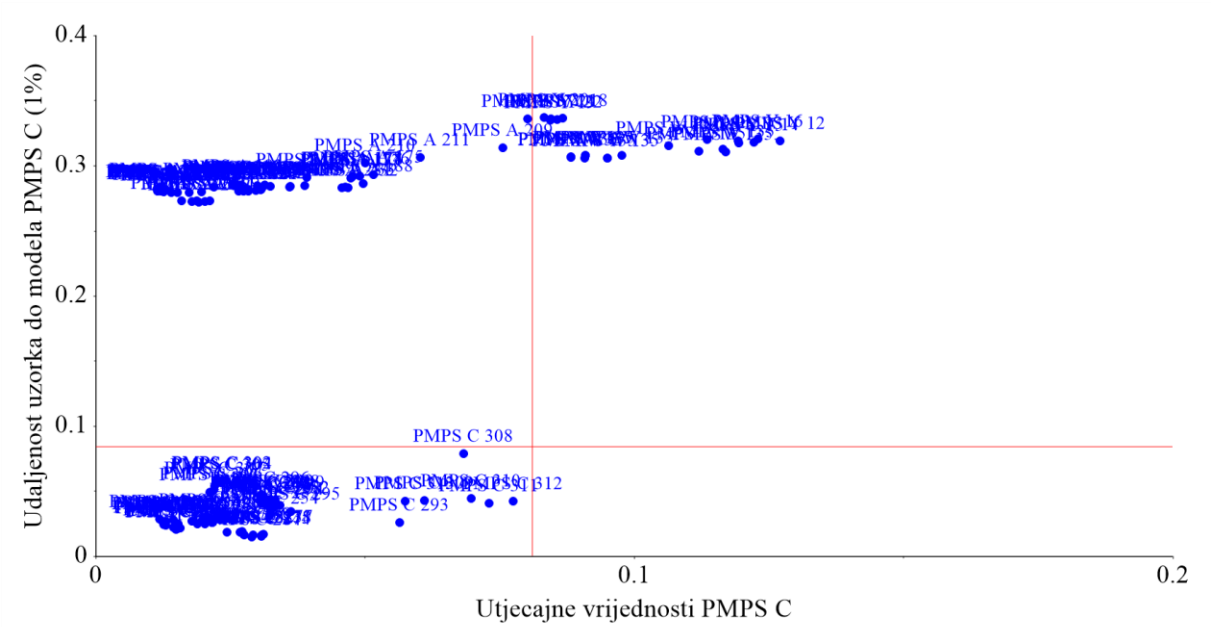


Slika 103. Cooman dijagram s pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).

Iz ovog Cooman dijagrama jasno se vidi da su svi uzorci nedvosmisleno dodijeljeni u odgovarajuće serogrupe (PMPS A, plavo; ili PMPS C, crveno) i da nema uzoraka između dvije serogrupe, što potvrđuje učinkovitost NIR SIMCA modela. Uzorci PMPS W135 i PMPS Y (zeleno), koji su korišteni kao negativne probe, očekivano nisu dodijeljeni niti u jednu klasu PMPS.



Slika 104. Udaljenosti uzorka PMPS A od modela (Si) s utjecajnim vrijednostima ovih uzoraka (Hi) za model PMPS A i pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).



Slika 105. Udaljenosti uzorka PMPS A od modela (Si) s utjecajnim vrijednostima ovih uzoraka (Hi) za model PMPS A i pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).

Prikaz udaljenosti uzoraka od modela (Si) s utjecajnim vrijednostima (Hi) za modele PMPS A i C sa postavljenim granicama za ove obje vrijednosti daje uvid u klasifikaciju PMPS u odgovarajući model. Na slikama 104. i 105. može se vidjeti da su svi (nepoznati) uzorci iz vanjskog validacijskog seta ispravno dodjeljeni u odgovarajuću serogrupu. Niti jedan uzorak nije kategoriziran u dvije serogrupe istovremeno niti je pogrešno identificiran kao pripadnik druge serogrupe, kao što je dodatno istaknuto u Tablici 6.

Tablica 6. Matrica zabune validacijskih parametara za NIR SIMCA model sa dva PC-a za svaku serogrupu i ukupni NIR SIMCA model, kod kojih su korišteni uzorci iz vanjskog test seta.

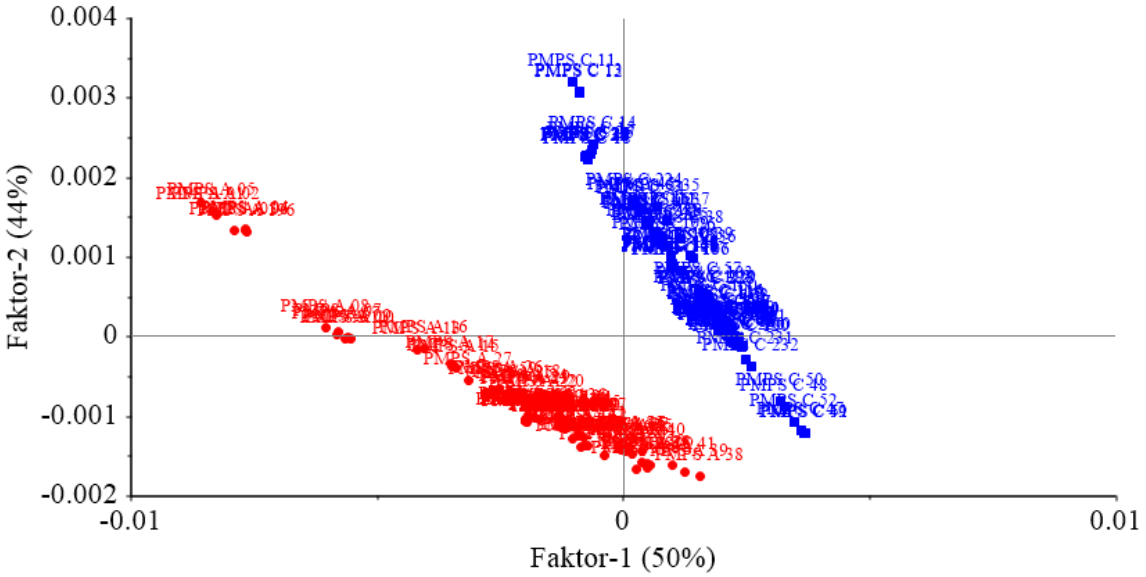
stvarno/predviđeno	klasa PMPS A	klasa PMPS C	nije klasificirano
klasa PMPS A	60	0	0
klasa PMPS C	0	62	0
klasa PMPS W135	0	0	12
klasa PMPS Y	0	0	12
CSNS	100%	100%	TSNS = 100%
CSPS	100%	100%	TSPS = 100%
CEFF	100%	100%	TEFF (2 PC) = 100%

CSNS, CSPS, CEFF, TSNS, TSPS i TEFF su opisani u poglavlju 2.5.3.2.

Temeljem svih ovdje opisanih rezultata može se zaključiti kako je primjena NIR spektroskopije u kombinaciji sa formiranim SIMCA modelom visoko učinkovita za identifikaciju PMPS A i PMPS C.

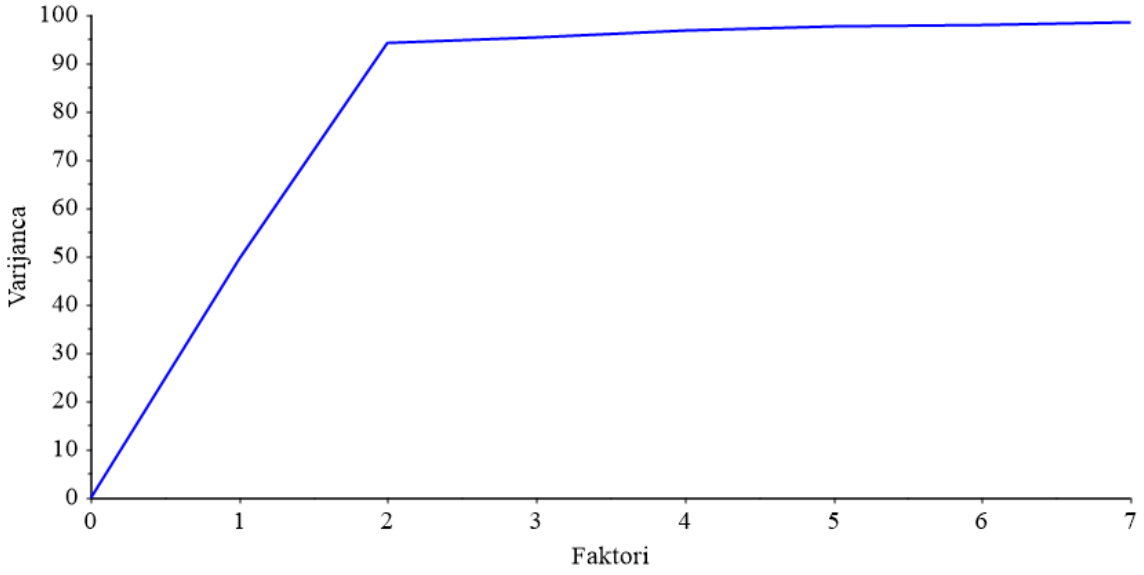
4.2.6. NIR PLS-DA model (Savitzky - Golay glačanje 3.9 s drugom derivacijom)

Uobičajeno, primjena PLS-DA modela za klasifikaciju dviju klasa (serogrupa) koristi dijagram faktorskih bodova kao prikaz klasificiranih uzoraka (Slika 106.).



Slika 106. Raspodjela faktorskih bodova matematički obrađenih (Savitzky-Golay glačanje 3.9 s drugom derivacijom) NIR spektara PMPS A i C u spektralnom području $\tilde{\nu}$ = 7768 - 3695 $cm^{-1}cm^{-1}$.

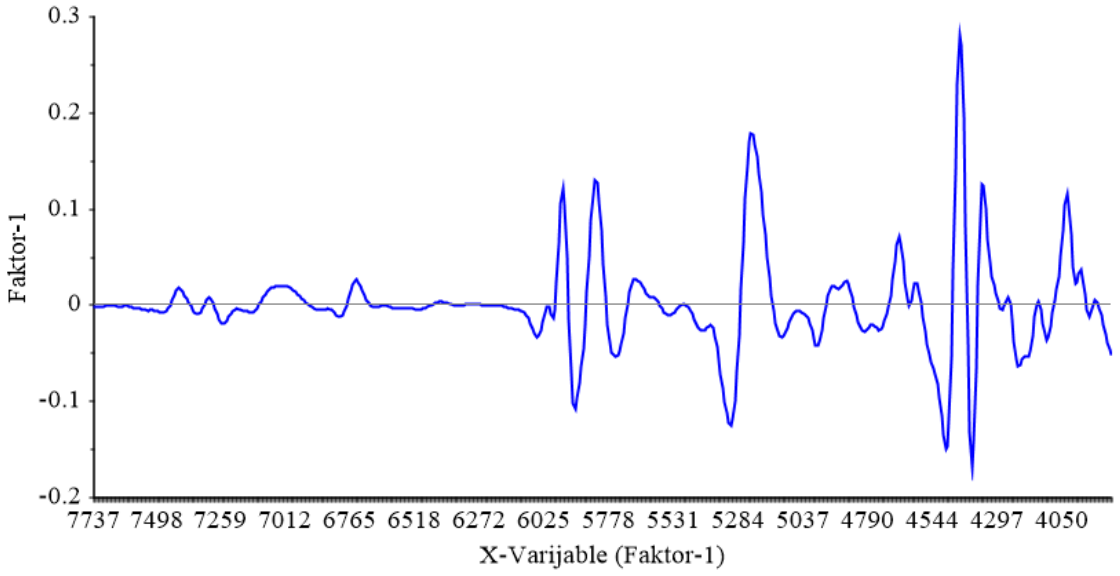
Na Slici 106. se vidi jasno razdvajanje među grupama PMPS A i PMPS C. Međutim, izvođenje zaključaka samo na temelju PLS dijagrama faktorskih bodova je nedostatno, jer na temelju ovoga dijagrama faktorskih bodova bez odgovarajuće validacije PLS-DA modela i to relevantnim test setom uzoraka je preoptimistično i može se doći do sasvim krivog zaključka. Kako bi odredili optimalni broj PLS faktora, potrebno je prikazati kumulativnu varijancu za svaki PLS faktor (Slika 107.).



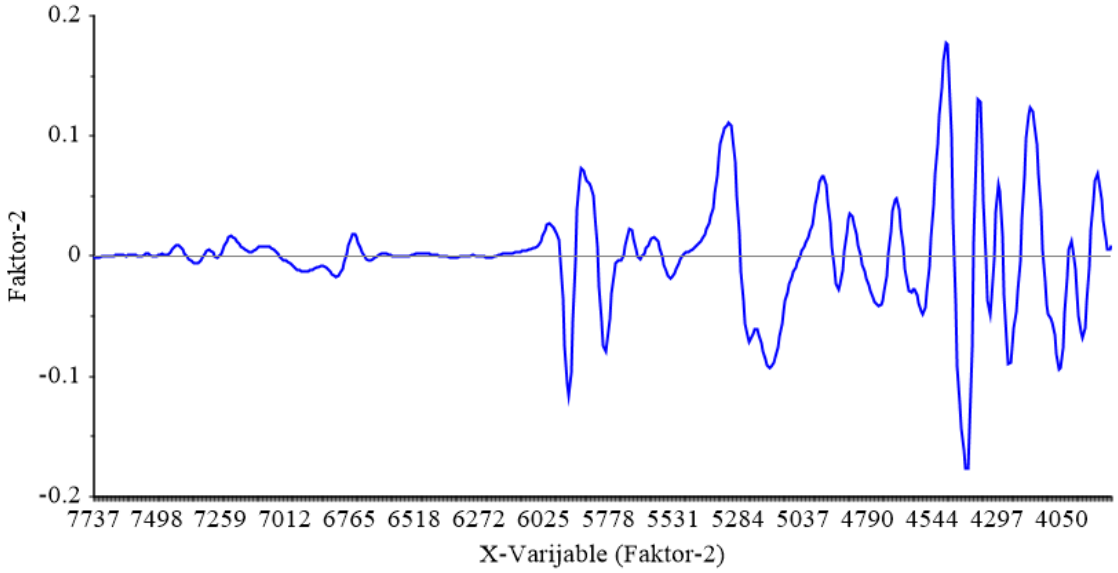
Slika 107. Kumulativna varijanca za svaki PLS faktor.

Na Slici 107. jasno se vidi da dva faktora obuhvaćaju 99 % kalibracijske varijance. Krivulja kalibracijske varijance ukazuje da su dva faktora optimalna za PLS model, jer nakon ovog faktora postoji samo neznatan porast u objašnjenoj varijanci. Ovaj mali broj latentnih varijabli (LV), odnosno PLS faktora, sugerira nisku korelaciju u NIR spektrima različitih klasa (serogrupa), ali i sličnosti u NIR spektrima unutar jedne klase (serogrupe).

Kako bi odredili najznačajnije valne brojeve na PLS faktore, bilo je potrebno načiniti dijagrame opterećenja (Slika 108 - 109.).



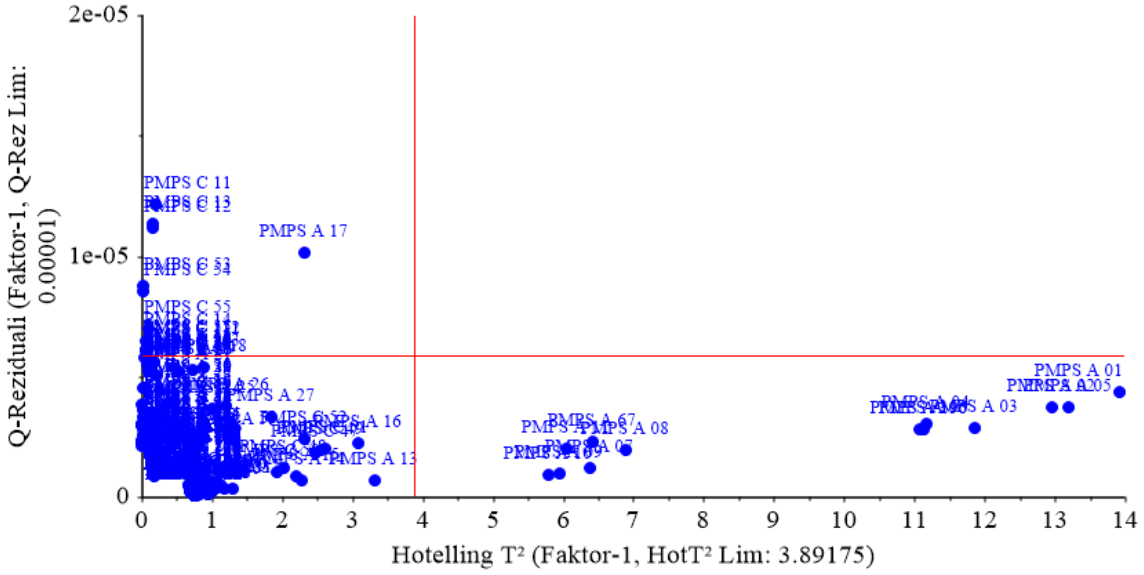
Slika 108. Profil opterećenja za prvu latentnu varijablu (LV).



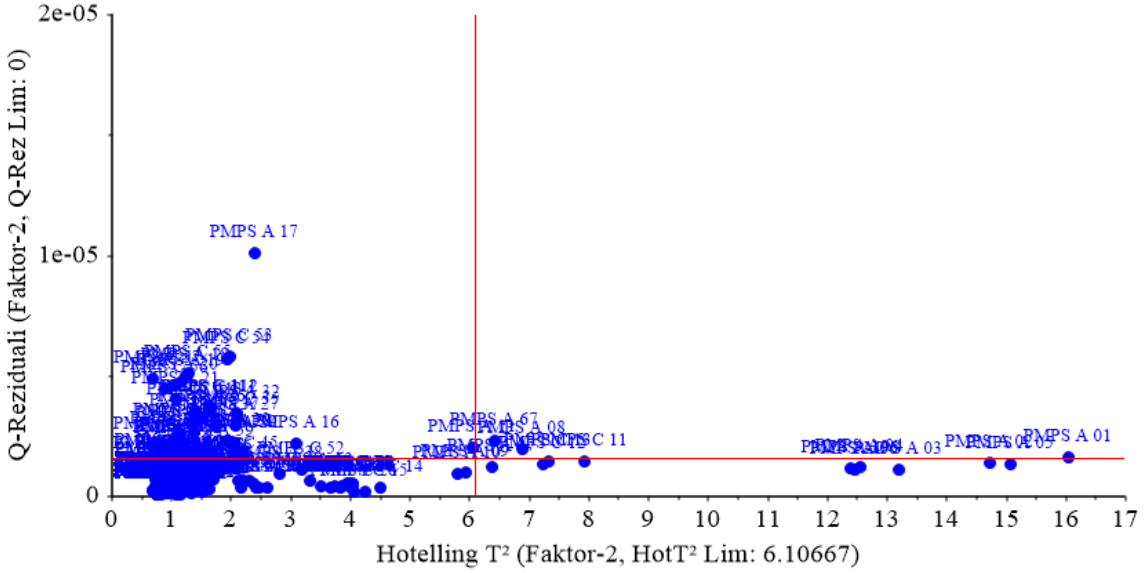
Slika 109. Profil opterećenja za drugu latentnu varijablu (LV).

Na slikama 108 i 109. se vidi da su valni brojevi u spektralnim područjima, $\tilde{\nu}$ = 5970 -5910 cm^{-1} i $\tilde{\nu}$ = 5780 - 5840 cm^{-1}, zatim $\tilde{\nu}$ = 5300 - 5100 cm^{-1}, $\tilde{\nu}$ = 5150 - 5240 cm^{-1} , $\tilde{\nu}$ = 4360 - 4450 cm^{-1} i $\tilde{\nu}$ = 4060 - 4150 cm^{-1} varijable koje su najviše odgovorne za razdvanje dviju klasa (serogrupa). Opisana spektralna regija $\tilde{\nu}$ = 5970 - 5910 cm^{-1} proizlazi od prvog višeg tona acetamide metil C-H asimetričnog istezanja; i $\tilde{\nu}$ = 5780 - 5840 cm^{-1} nastaje iz prvog višeg tona metilen C-H asimetričnog istezanja; spektralna regija $\tilde{\nu}$ = 5300 - 5100 cm^{-1} odgovara O-H kombinacijskim vibracijama; spektralno područje $\tilde{\nu}$ = 5150 - 5240 cm^{-1} odgovara kombinaciji polisaharidnog O-H istezanja, H-O-H deformacije i O-H savijanja, spektralno područje $\tilde{\nu}$ = 4360 - 4450 cm^{-1} odgovara kombinaciji C–H istezanja i CH_2 deformacije a spektralno područje $\tilde{\nu}$ = 4060 - 4150 cm^{-1} odgovara kombinaciji istezanja C-H, C-C i C-O-C (Workman, 2001).

Kako bi provjerili prisutnost netipičnih, odnosno ekstremnih uzoraka, načinjeni su Q reziduali i Hotelling T^2, za obe latentne varijable (Slike 110. - 111.).



Slika 110. Hotelling T^2 statistika i Q-reziduali za prvi PLS faktor s pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).

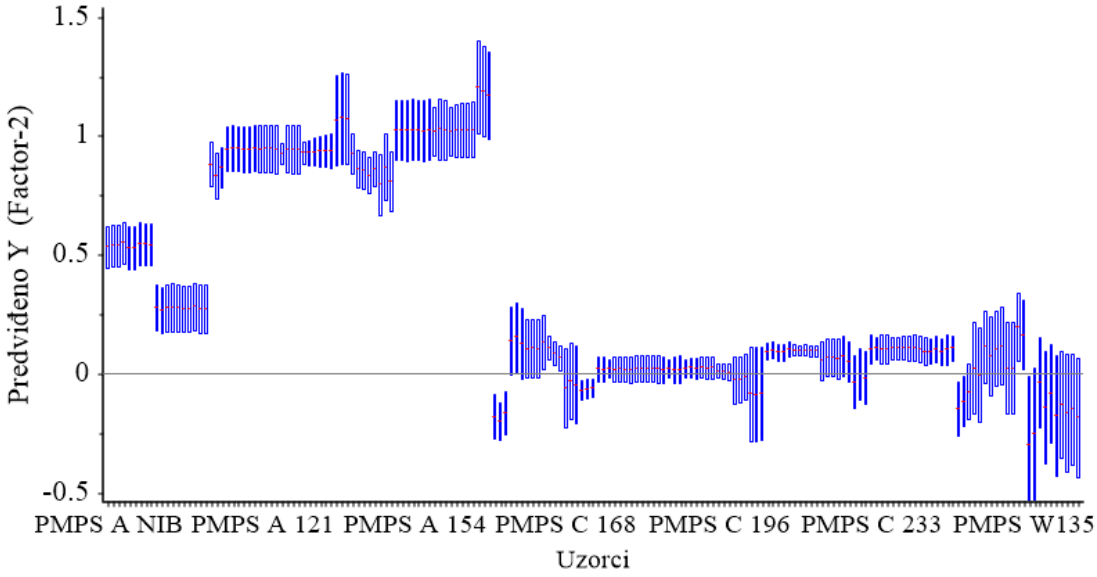


Slika 111. Hotelling T^2 statistika i Q-reziduali za drugi PLS faktor s pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).

Iz Slika 110. i 111. jasno se vidi prisutnost ekstremnih uzoraka, ali nema uzoraka koji bi predstavljali opasnost za model, odnosno netipičnih uzoraka.

4.2.6.1. Optimizacija NIR PLS-DA modela

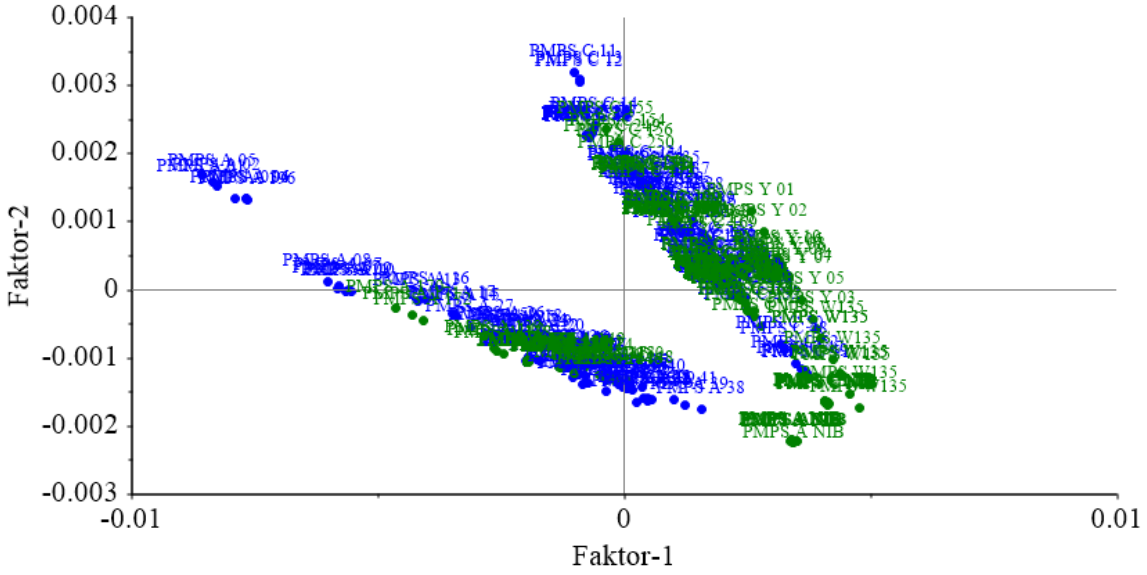
Kako bi optimirali NIR PLS-DA model s dvije latentne varijable (LV), provedena je identifikacija uzoraka iz test seta 1 s pomoću ovoga formiranoga modela. Rezultati ove identifikacije prikazani su na Slici 112.



Slika 112. Predviđene vrijednosti uzoraka test seta 1 s procjenjenim odstupanjem dobivene formiranim NIR PLS-DA modelom sa dvije latentne varijable.

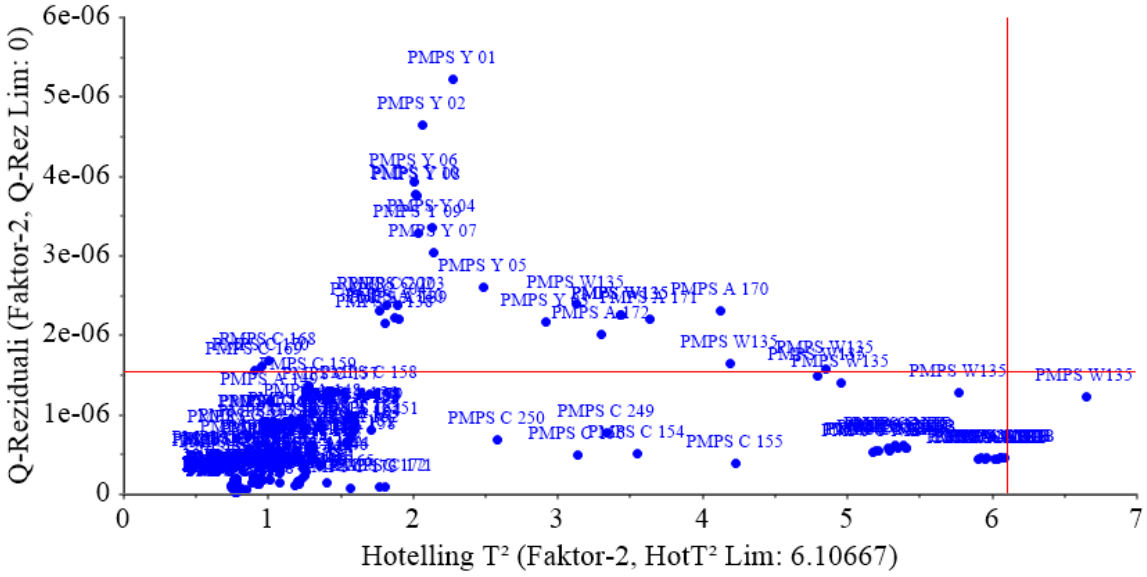
Slika 112. prikazuje kako PLS-DA model sa dva PLS faktora identificira nepoznate uzorke iz test seta 1. Na ovoj se slici vidi da su svi uzorci PMPS C ispravno dodjeljeni odgovarajućoj serogrupi. Nadalje, uzorci PMPS A su oko idealne vrijednosti 1 i ovi uzorci su ispravno identificirani i dodijeljeni u odgovarajuću serogrupu - PMPS A. Također su svi uzorci PMPS W135 i Y pridruženi klasi C, što je i bilo za očekivati i to zbog sličnosti u njihovoj kemijskoj strukturi (PMPS C, W135 i Y). Iz dobivenih rezultata može se zaključiti da NIR PLS-DA model s dva PLS faktora ima vrlo dobru sposobnost identifikacije PMPS A i C.

Kako bi prikazali preklapanje uzoraka iz test seta 1 i trening seta, načinjen je dijagram raspodjele faktorskih bodova za ova dva seta uzoraka.



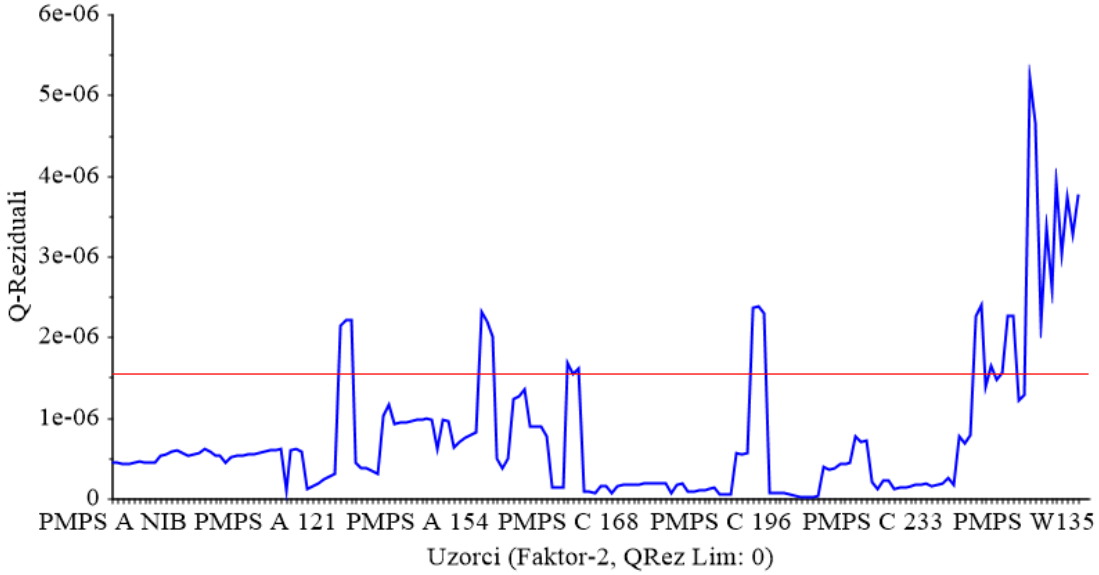
Slika 113. Raspodjela faktorskih bodova matematički obrađenih (Savitzky-Golay glačanje 3.9 s drugom derivacijom) NIR spektara trening seta i test seta 1 PMPS A i C u području $\tilde{\nu}$ = 7768 - 3695 cm^{-1}. Uzorci trening seta prikazani su plavom, a test seta 1 zelenom bojom.

Slika 113. prikazuje jasno odvajanje među uzorcima PMPS A i PMPS C te pridruživanje uzoraka PMPS W135 i PMPS Y uzorcima PMPS C. Uzorci test seta 1 preklapaju se sa uzorcima trening seta ispravno sa klasom meningokoknih polisaharida odgovarajuće serogrupe. Obzirom da su uzorci PMPS C, W135 i Y dodijeljeni kao pripadnici klase PMPS C, bilo je potrebno ispitati dali će se dodatnim statističkim analizama moći identificirati kao netipični uzorci. Prisutnost netipičnih uzoraka u PLS-DA modelu može se procjeniti pomoću Hotelling T^2 statistike i Q-reziduala (Slika 114.). Budući da PLS-DA model pruža aproksimativno rješenje za identifikaciju uzoraka PMPS, Q-statistika se koristi za procjenu usklađenosti svakog uzorka s formiranim modelom. Ako su vrijednosti Q visoke za pojedini uzorak to znači da dotični uzorak ima veliku varijaciju izvan modela i nije u skladu s modelom. Kada su vrijednosti Hotelling T^2 statistike visoke za pojedini uzorak, ukazuju na to da taj uzorak ima veliku varijaciju unutar modela.



Slika 114. Hotelling T^2 statistika i Q-reziduali s pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).

Na Slici 114. se vidi da uzorci PMPS W135 i Y imaju visoke Hotelling T^2 vrijednosti i visoke vrijednosti Q-reziduala, što je potpuno u skladu s očekivanim rezultatom klasifikacije ovih PMPS. Ovi uzorci PMPS prepoznati su kao netipični uzorci.



Slika 115. Q reziduali uzoraka test seta 1 s pripadajućom graničnom linijom (crvena linija).

Slike 114. i 115. zapravo potvrđuju da uzorci PMPS W135 i Y imaju visoke Hotelling T^2 vrijednosti i da uzorci PMPS W135 i Y imaju Q-reziduale više od granične vrijednosti (crvena linija, Slika 115.). Može se zaključiti da su PMPS W135 i Y netipični uzorci, kao što je prethodno zaključeno i svakako očekivano kod planiranja ovih eksperimenata.

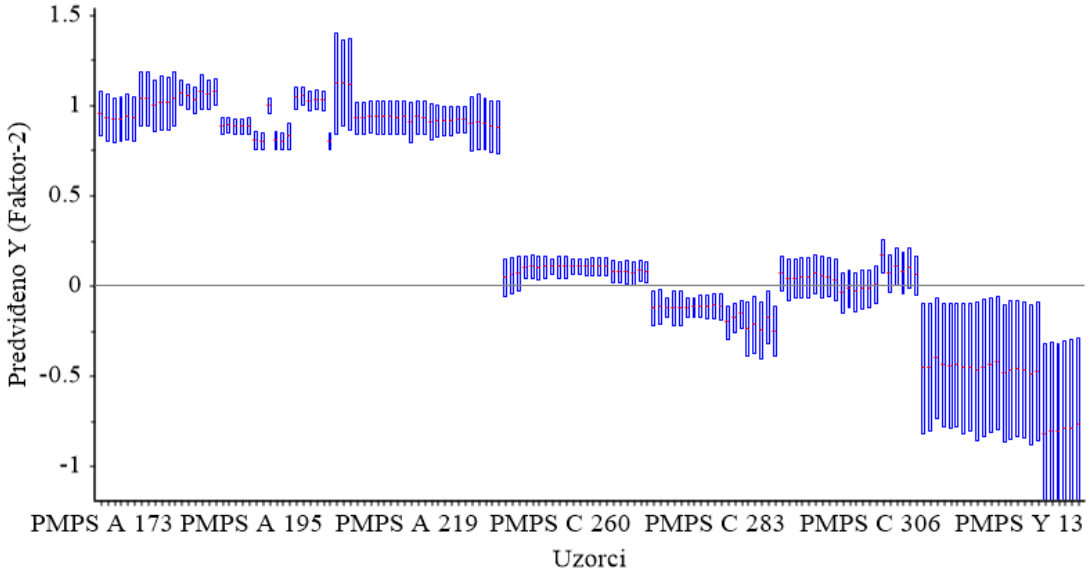
Tablica 7. Matrica zabune validacijskih parametara za PMPS uzorke iz test seta 1 dobivenih NIR PLS-DA modelom s dva PLS faktora.

stvarno/predviđeno	klasa PMPS A	klasa PMPS C	nije klasificirano
klasa PMPS A	61	0	0
klasa PMPS C	0	98	0
klasa PMPS W135	0	10	0
klasa PMPS Y	0	10	0
CSNS	100%	100%	TSNS = 100%
CSPS	100%	0%	TSPS = 0%
CEFF	100%	0%	TEFF (2 PLS) = 0%

CSNS, CSPS, CEFF, TSNS, TSPS i TEFF su opisani u poglavlju 2.5.3.2.

4.2.6.2. Validacija NIR PLS-DA modela

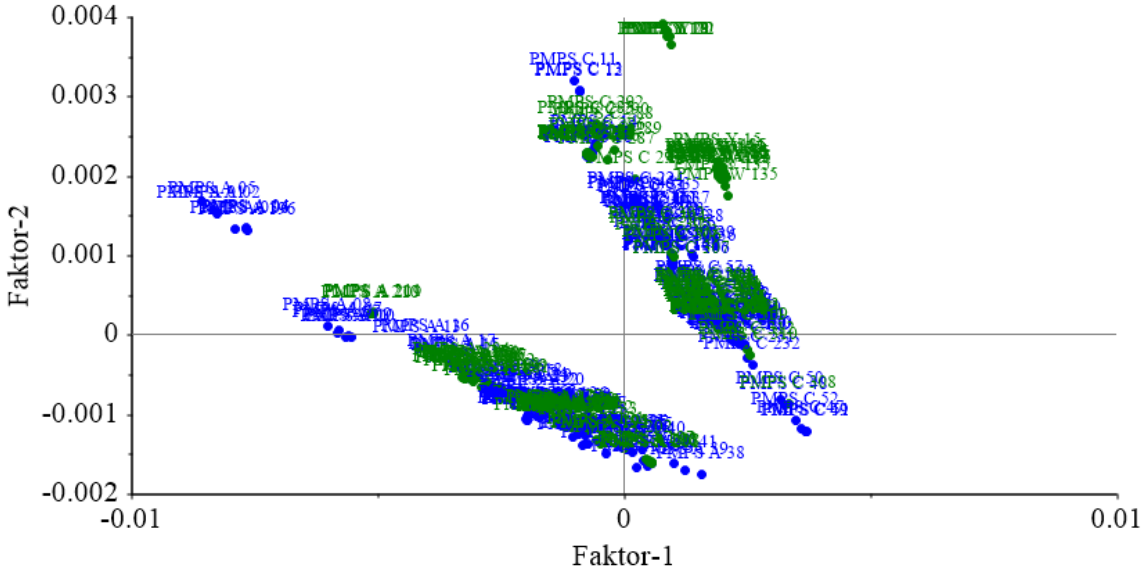
Kako bi validirali NIR PLS-DA model s dva PLS faktora, provedena je identifikacija nepoznatih uzoraka PMPS iz vanjskog test seta. Rezultati identifikacije ovim PLS-DA modelom prikazani su predviđenim vrijednostima (Slika 116.).



Slika 116. Predviđene vrijednosti uzoraka PMPS vanjskog test seta s procjenjenim odstupanjem dobivene fromiranim NIR PLS-DA modelom s dvije latentne varijable.

Na slici 116. identificirani su uzorci PMPS iz vanjskog test seta formiranim NIR PLS-DA modelom s dva PLS faktora. Ova slika jasno prikazuje da su svi uzorci PMPS A i PMPS C ispravno klasificirani (dodjeljeni odgovarajućoj serogrupi), dok su uzorci PMPS W135 i PMPS Y, koji su korišteni kao negativna proba, očekivano dodijeljeni klasi PMPS C. Sličan rezultat dobiven je kod optimizacije NIR PLS-DA modela pomoću test seta 1 i to zbog sličnosti kemijske strukture PMPS W135 i PMPS Y s PMPS C. Velike nesigurnosti ukazuju na to da klasifikacija PLS-DA modelom nije u cjelosti pouzdana, što se jasno vidi na ovoj slici kod uzoraka PMPS W135 i PMPS Y. Dobiveni rezultati ukazuju na visoku efikasnost klasifikacije formiranim NIR PLS-DA modelom s dva PLS faktora.

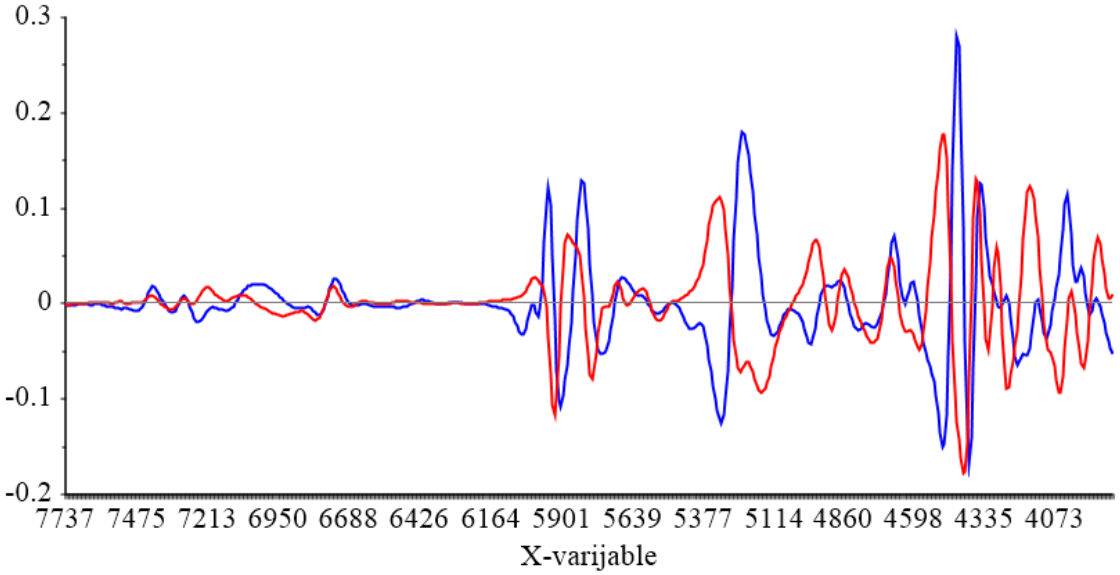
Kako bi utvrdili preklapanje među uzorcima PMPS iz kalibracijskog seta i vanjskog test seta i njihovu raspodjelu, potrebno je prikazati raspodjelu faktorskih bodova (Slika 117.).



Slika 117. Raspodjela faktorskih bodova matematički obrađenih (Savitzky-Golay glačanje 3.9 s drugom derivacijom) NIR spektara trening seta i vanjskog validacijskog seta PMPS A i C u području $\tilde{\nu}$ = 7768 - 3695 cm^{-1}. Uzorci trening seta prikazani su plavom, a test seta 1 zelenom bojom.

Raspodjela faktorskih bodova (Slika 117.) jasno prikazuje odvajanje PMPS A i PMPS C. Uzorci PMPS W135 i Y, koji su korišteni kao negativna proba, identificirani su NIR PLS-DA modelom kao PMPS C, zbog sličnosti u njihovoj kemijskoj strukturi sa ovim meningokoknim polisaharidom.

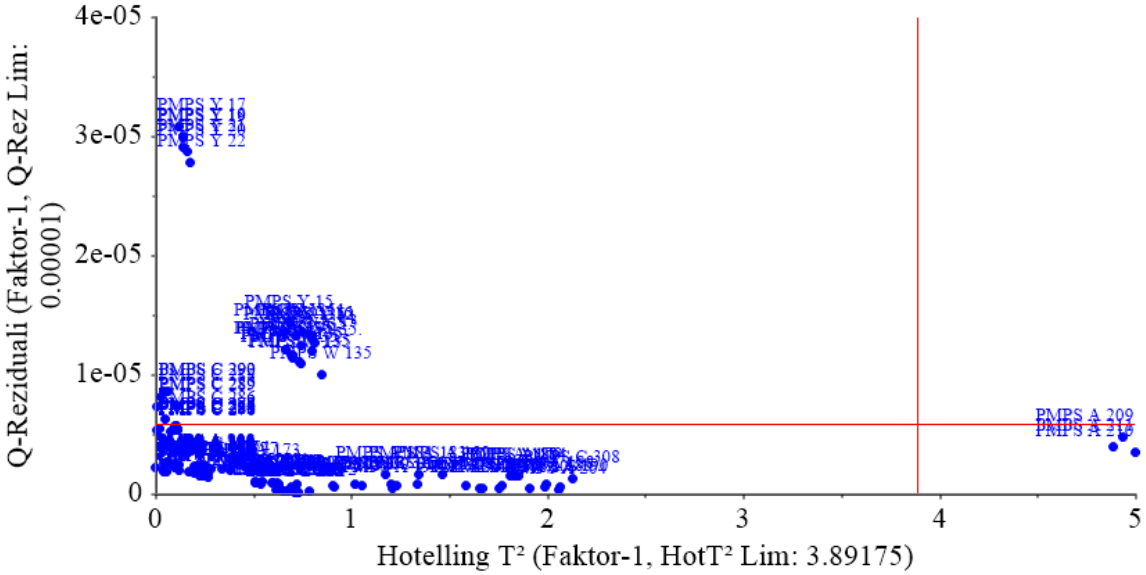
Kako bi prikazali valne brojeve s najvećim utjecajem na latentne varijable ,a time i valne brojeve odgovorne za međusobno razdvajanje polisaharida A i C, prikazano je faktorsko opterećenje (Slika 118.).



Slika 118. Profil opterećenja za prve dvije latentne varijable po valnim brojevima.

Na slici faktorskog opterećenja za prve dvije latentne varijable (Slika 118.) može se vidjeti da su za razdvanje dvaju PMPS najviše odgovorne ove spektralne regije: $\tilde{\nu}$ = 5970 - 5910 cm^{-1} koje proizlazi od prvog višeg tona acetamid metil C–H asimetričnog istezanja; $\tilde{\nu}$ = 5780 - 5840 cm^{-1} proizlazi od prvog višeg tona metilen C–H asimetričnog istezanja; spektralno područje $\tilde{\nu}$ = 5300 - 5100 cm^{-1} odgovara O-H kombinacijskim vibracijama; spektralno područje $\tilde{\nu}$ = 5150 - 5240 cm^{-1} odgovara kombinaciji polisaharidnog O-H istezanja, H-O-H deformacije i O-H savijanja spektralno područje $\tilde{\nu}$ = 4360 - 4450 cm^{-1} odgovara kombinaciji C–H istezanja i CH_2 deformacije regija, dok $\tilde{\nu}$ = 4060 - 4150 cm^{-1} odgovara kombinaciji istezanja C- H, C-C istezanja i C-O-C istezanja (Workman, 2001).

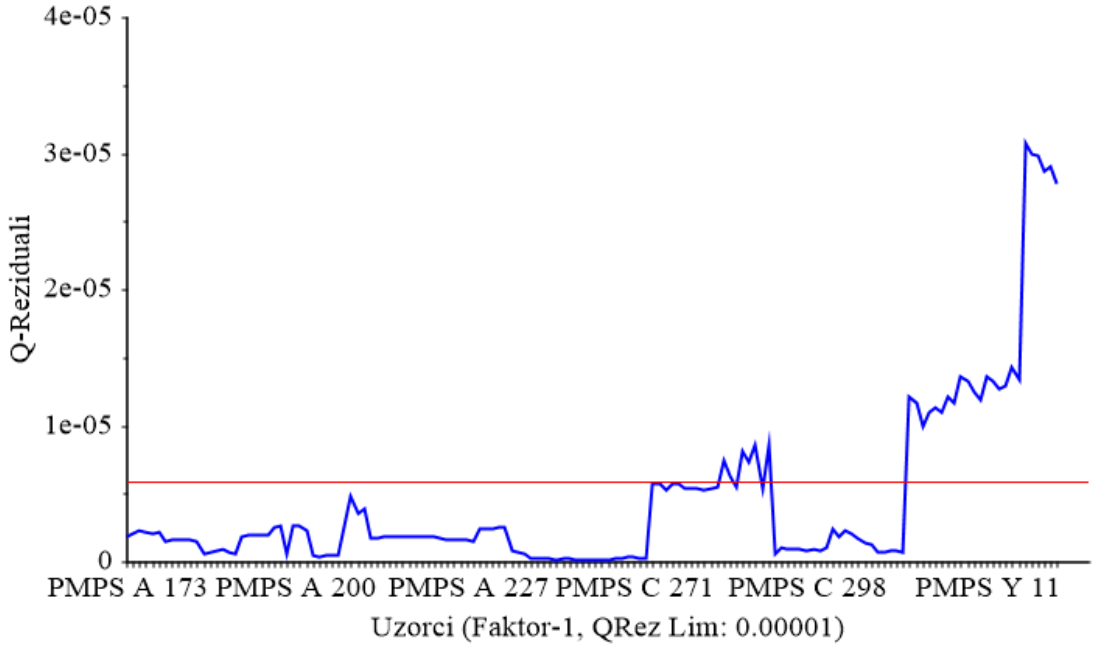
Kako bi se identificirali netipični PMPS uzorci,odnosno, kako bi potvrdili da su uzorci PMPS W135 i PMPS Y identificirani kao netipični uzorci, i znatno različiti od ostatka uzoraka PMPS C čijoj su klasi dodijeljeni, načinjena je Hotelling T^2 statistika i Q-reziduali (Slike 119. - 121.).



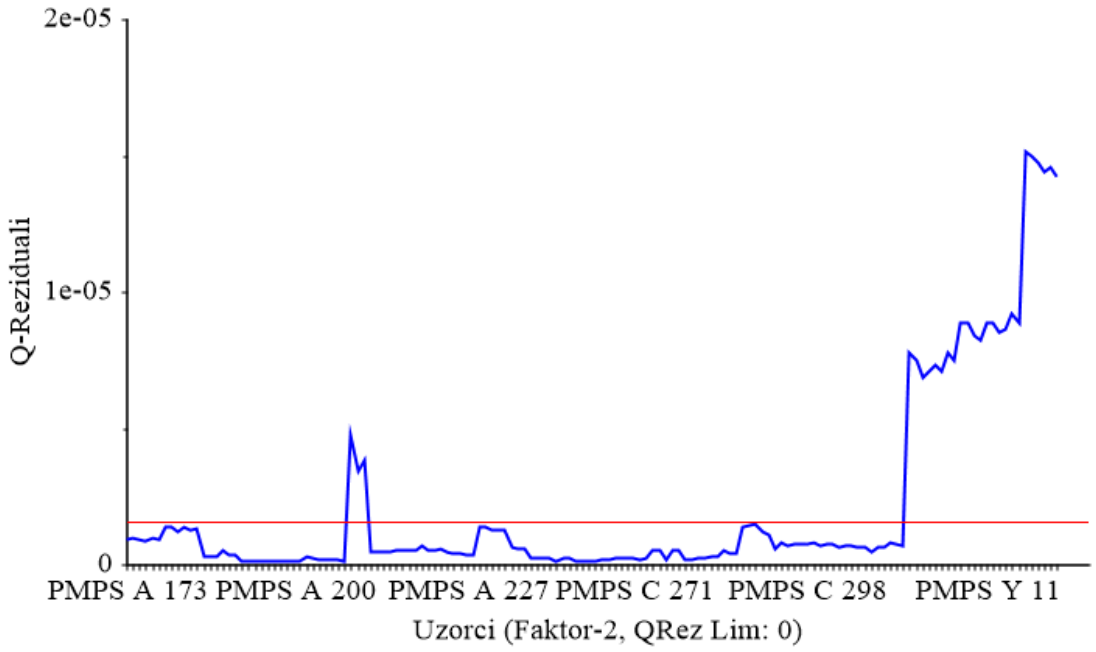
Slika 119. Hotelling T^2 statistika i Q-reziduali s pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).

Slika 119. prikazuje Hotelling T^2 statistiku na osi apscisa s odgovarajućim kritičnim granicama (okomita crvena linija), a Q-residuali na osi ordinata također s kritičnim granicama (horizontalna crvena linija). Hotelling T^2 statistika opisuje udaljenost pojedinog uzorka od centra PLS-DA modela u rasponu latentnih varijabli, a Q-statistika je suma kvadrata reziduala svih varijabli za svaki pojedini PMPS uzorak.

Slike 119 - 121. jasno prikazuju kako su uzorci PMPS W135 i PMPS Y identificirani kao netipični uzorci visokih Q reziduala.



Slika 120. Q-reziduali uzoraka PMPS vanjskog test seta za jedan PLS faktor s pripadajućom graničnom linijom (crvena linija).



Slika 121. Q-reziduali uzoraka PMPS vanjskog test seta za dva PLS faktora s pripadajućom graničnom linijom (crvena linija).

4.2.6.2.1. Validacijski parametri NIR PLS-DA modela

Validacijski parametri (FoM, Tablica 8.) se koriste za procjenu prediktivne sposobnosti PMPS uzoraka iz vanjskog seta pomoću NIR PLS-DA modela sa dva PLS faktora.

Tablica 8. Matrica zabune validacijskih parametara za PMPS uzorke iz vanjskog validacijskog seta dobivenih NIR PLS-DA modelom s dva PLS faktora.

stvarno/predviđeno	klasa PMPS A	klasa PMPS C	nije klasificirano
klasa PMPS A	60	0	0
klasa PMPS C	0	62	0
klasa PMPS W135	0	12	0
klasa PMPS Y	0	12	0
CSNS	100%	100%	TSNS = 100%
CSPS	100%	0%	TSPS = 0%
CEFF	100%	0%	TEFF (2 PLS) = 0%

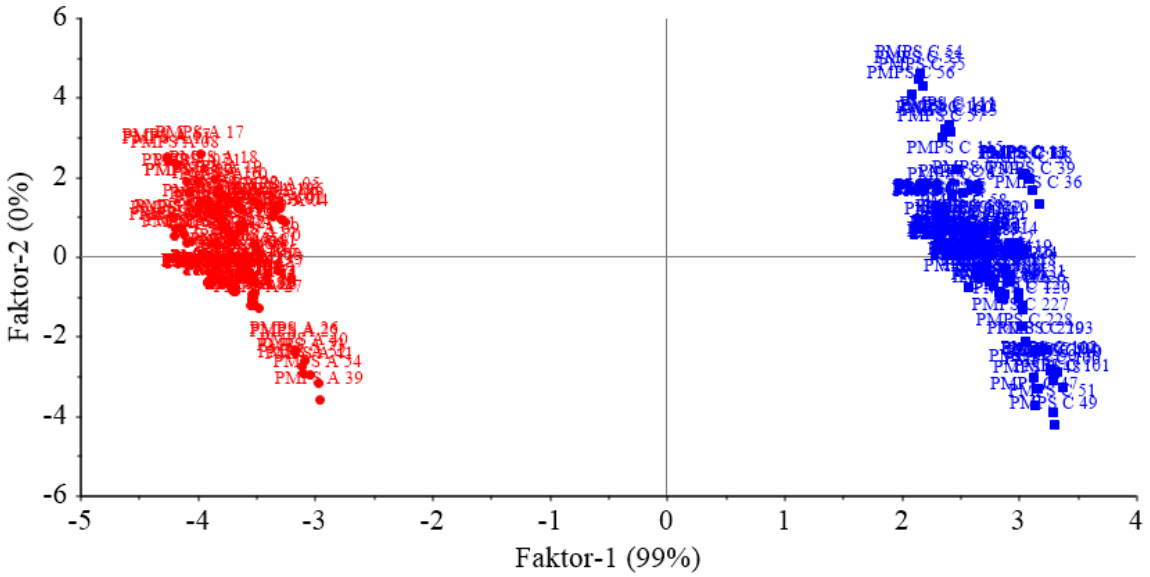
CSNS, CSPS, CEFF, TSNS, TSPS i TEFF su opisani u poglavlju 2.5.3.2.

Iz dobivenih rezultata može se zaključiti da je formirani NIR PLS-DA model izuzetno učinkovit za identifikaciju nepoznatih uzoraka PMPS A i C.

Kako bi usporedili rezultate dobivene klasifikacijskim modelima formiranih nakon različitih matematičkih predobrada, također su prikazani rezultati validacije NIR PLS-DA model matematički obrađenih (Savitzky-Golay glačanje 3.9 s drugom derivacijom i SNV) NIR spektralnih podataka.

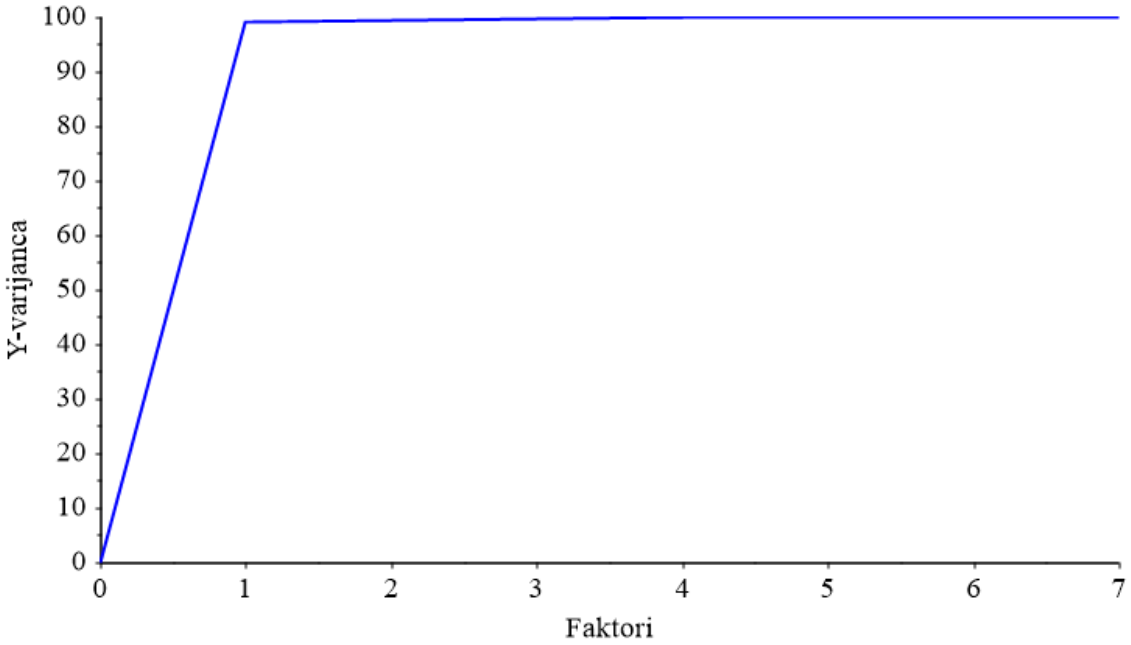
4.2.7. NIR PLS-DA model (Savitzky - Golay glačanje 3.9 s drugom derivacijom i SNV)

Klasifikacija uzoraka primjenom PLS-DA modela za klasifikaciju dviju klasa (serogrupa) kao prikaz koristi dijagram faktorskih bodova (Slika 122.).



Slika 122. Raspodjela faktorskih bodova matematički obrađenih (Savitzky-Golay glačanje 3.9 s drugom derivacijom i SNV-om) NIR spektara PMPS A i C u spektralnom području $\tilde{\nu}$ = 7768 - 3695 $cm^{-1}cm^{-1}$.

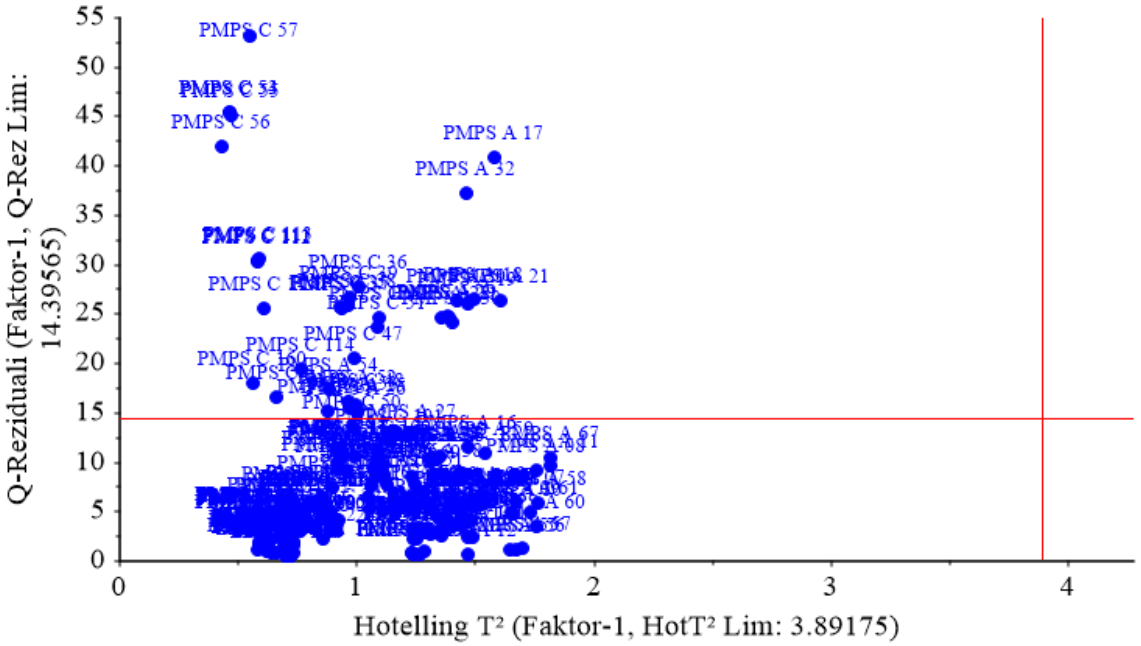
Na Slici 122. se vidi jasno razdvajanje među grupama PMPS A i PMPS C. Međutim, kako je već rećeno (poglavlje ispred) izvođenje zaključaka samo na temelju PLS dijagrama faktorskih bodova je nedostatno, te je provedena odgovarajuća validacija PLS-DA modela.
Kako bi odredili optimalni broj PLS faktora, potrebno je prikazati kumulativnu varijancu za svaki PLS faktor (Slika 123.).



Slika 123. Kumulativna varijanca za svaki PLS faktor.

Na Slici 123. jasno se vidi da jedan PLS faktor obuhvaća 99 % kalibracijske varijance. Kalibracijska krivulja ukazuje da je jedan faktor optimalan za PLS model, jer nakon ovog faktora postoji samo neznatan porast u objašnjenoj varijanci.

Kako bi se identificirali netipični PMPS uzorci, načinjena je Hotelling T^2 statistika i Q-reziduali (Slika 124.).

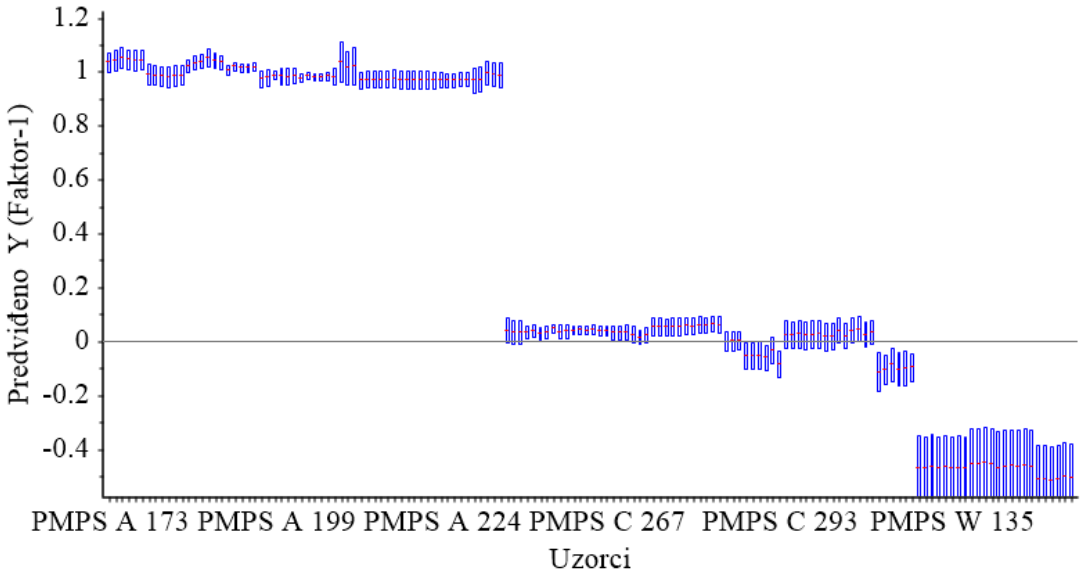


Slika 124. Hotelling T^2 statistika i Q-reziduali s pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).

Iz Slike 124. jasno se vidi prisutnost ekstremnih uzoraka. Nije prisutan niti jedan netipičan uzorak koji bi predstavljao opasnost za model.

4.2.7.1. Validacija NIR PLS-DA modela

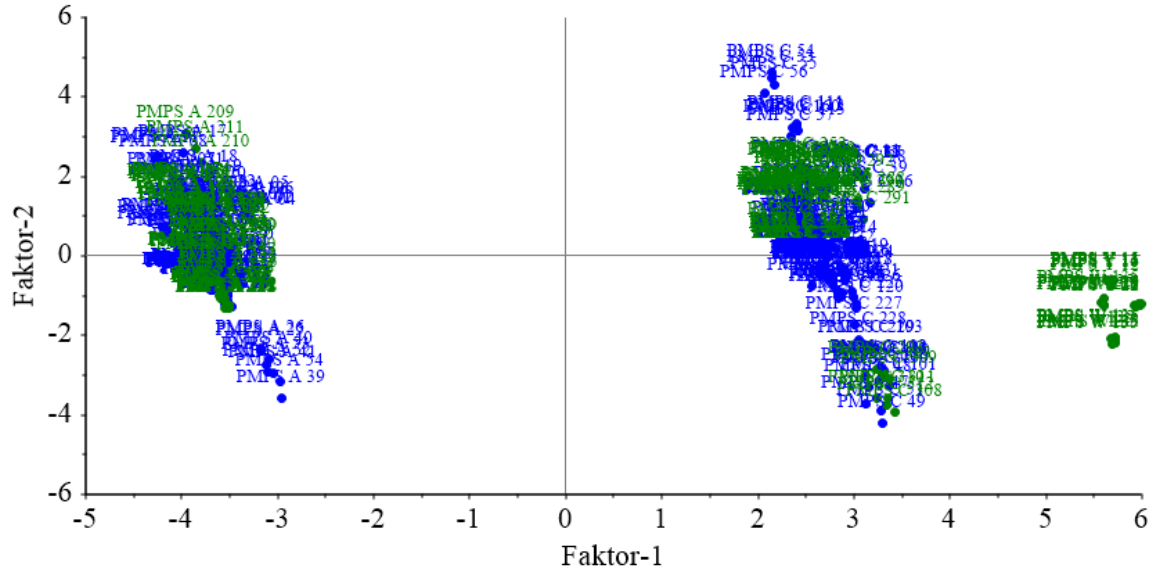
Kako bi validirali NIR PLS-DA model s jednim PLS faktorom provedena je identifikacija uzoraka iz vanjskog test seta s pomoću ovoga formiranoga modela. Rezultati ove identifikacije prikazani su na Slici 125.



Slika 125. Predviđene vrijednosti uzoraka vanjskog test seta s procjenjenim odstupanjem dobivene formiranim NIR PLS-DA modelom s jednim PLS faktorom.

Na slici 125. identificirani su uzorci PMPS iz vanjskog test seta formiranim NIR PLS-DA modelom s jednim PLS faktorom. Ova slika jasno prikazuje da su svi uzorci PMPS A i PMPS C ispravno klasificirani (dodjeljeni odgovarajućoj serogrupi), dok su uzorci PMPS W135 i PMPS Y, koji su korišteni kao negativna proba, očekivano dodijeljeni klasi PMPS C. Velike nesigurnosti ukazuju na to da klasifikacija PLS-DA modelom nije u cjelosti pouzdana, što se jasno vidi na ovoj slici kod uzoraka PMPS W135 i PMPS Y. Dobiveni rezultati ukazuju na visoku efikasnost klasifikacije formiranim NIR PLS-DA modelom s jednim PLS faktorom.

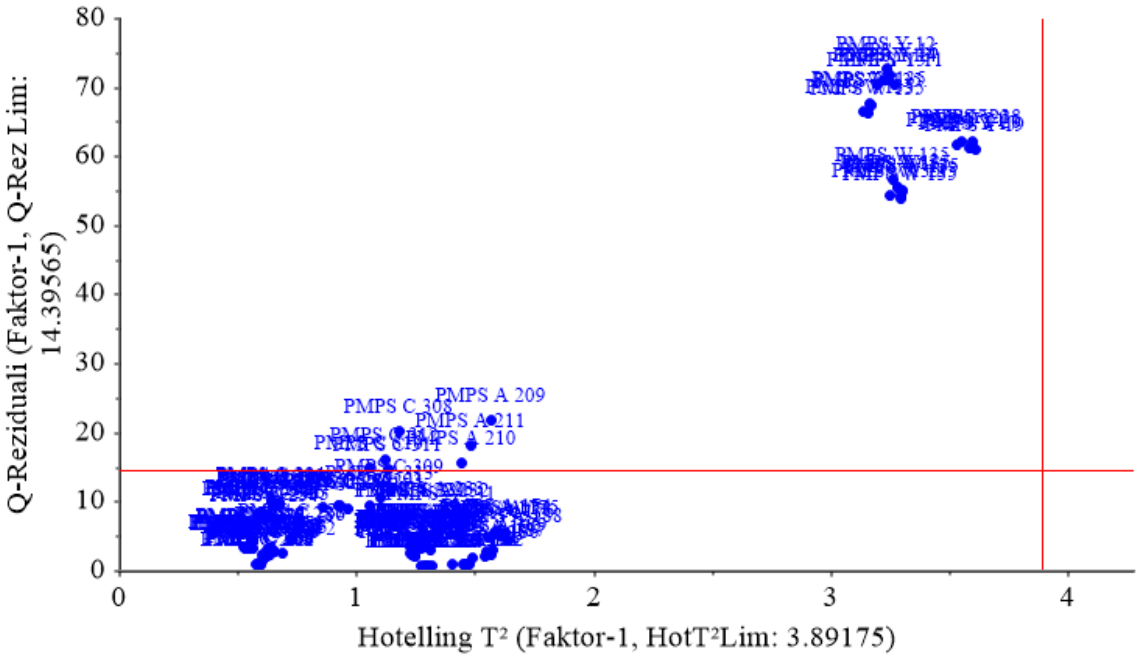
Kako bi utvrdili preklapanje među uzorcima PMPS iz kalibracijskog seta i vanjskog test seta i njihovu raspodjelu, potrebno je prikazati raspodjelu faktorskih bodova (Slika 126.).



Slika 126. Raspodjela faktorskih bodova matematički obrađenih (Savitzky-Golay glačanje 3.9 s drugom derivacijom i SNV) NIR spektara trening seta i vanjskog validacijskog seta PMPS A i C u području $\tilde{\nu}$ = 7768 - 3695 cm^{-1}. Uzorci trening seta prikazani su plavom, a vanjskog test seta zelenom bojom.

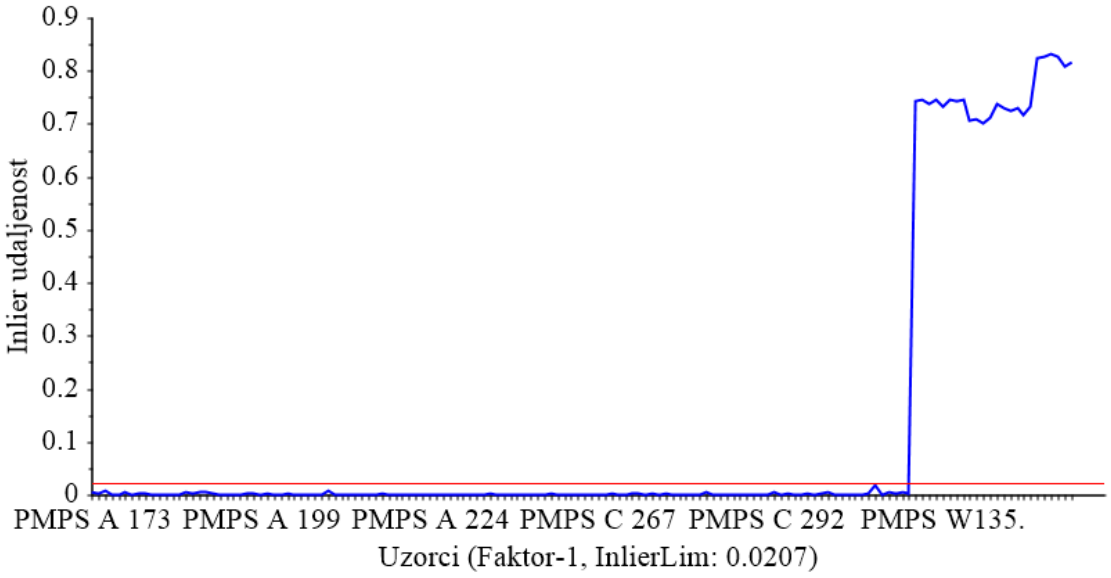
Raspodjela faktorskih bodova (Slika 126.) jasno prikazuje odvajanje PMPS A i PMPS C. Uzorci PMPS W135 i Y, koji su korišteni kao negativna proba, identificirani su NIR PLS-DA modelom kao PMPS C, zbog sličnosti u njihovoj kemijskoj strukturi sa ovim meningokoknim polisaharidom.

Kako bi se identificirali netipični PMPS uzorci, načinjena je prikaz Hotelling T^2 statistike i Q-reziduala (Slika 127.).

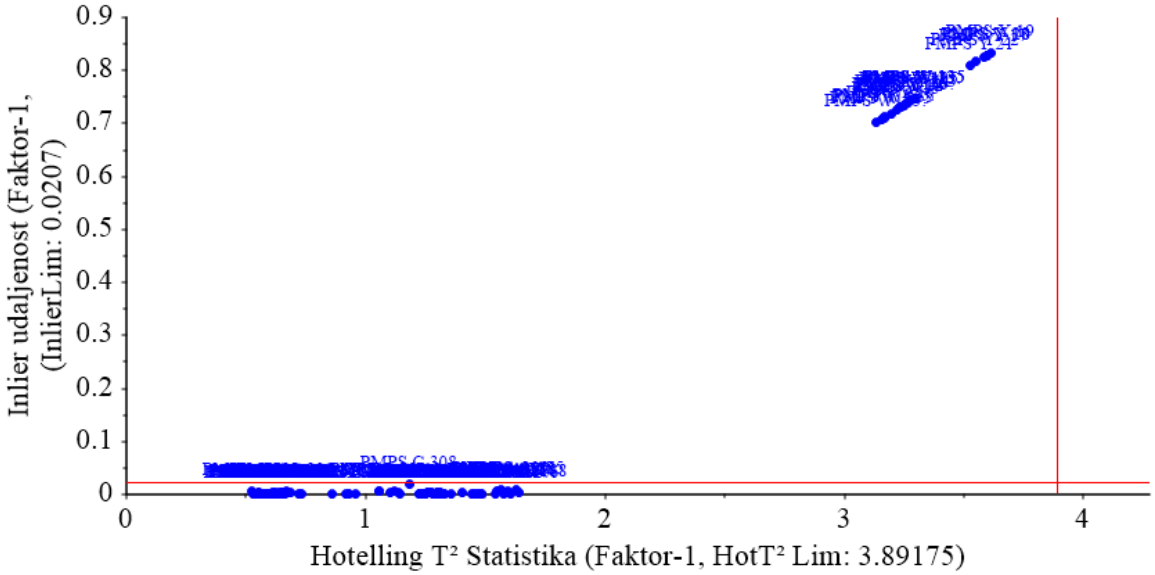


Slika 127. Hotelling T^2 statistika i Q-reziduali s pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).

Slika 127. jasno prikazuje kako su uzorci PMPS W135 i PMPS Y identificirani kao netipični uzorci visokih Q reziduala i visoke Hotelling T^2.



Slika 128. "Inlier" udaljenost s pripadajućom graničnom linijom (crvena linija).



Slika 129. "Inlier" i Hotelling T^2 udaljenosti s pripadajućim graničnim linijama (okomita i vodoravna crvena linija).

Da bi identifikacija uzoraka pomoću PLS-DA modela bila pouzdana, uzorci iz vanjskog test seta ne smiju biti predaleko od uzoraka iz kalibracijskog seta. To se utvrđuje prikazom "Inlier" udaljenosti (Slike 128. i 129.). Projekcija uzorka u PLS-DA modelu također ne smije biti previše udaljena od središta ovog modela. To se provjerava pomoću Hotelling udaljenosti.
Uzorci koji se nalaze izvan graničnih linija, njihova identifikacija PLS-DA modelom nije pouzdana. Na Slici 100. se vidi da se svi uzorci iz vanjskog test seta osim uzoraka PMPS W135 I PMPS Y nalaze ispod granica udaljenosti "Inliera", što znači da su ti uzorci slični onima koji su se koristili tijekom formiranja ovog PLS-DA modela (uzorci iz kalibracijskog seta).

4.2.7.1.1. Validacijski parametri NIR PLS-DA modela

Validacijski parametri (FoM, Tablica 9.) se koriste za procjenu prediktivne sposobnosti NIR PLS-DA modela sa jednim PLS faktorom uzoraka PMPS iz vanjskog test seta.

Tablica 9. Matrica zabune validacijskih parametara za PMPS uzorke iz vanjskog validacijskog seta dobivenih NIR PLS-DA modelom s jednim PLS faktorom.

stvarno/predviđeno	klasa PMPS A	klasa PMPS C	nije klasificirano
klasa PMPS A	60	0	0
klasa PMPS C	0	62	0
klasa PMPS W135	0	12	0
klasa PMPS Y	0	12	0
CSNS	100%	100%	TSNS = 100%
CSPS	100%	0%	TSPS = 0%
CEFF	100%	0%	TEFF (1 PLS) = 0%

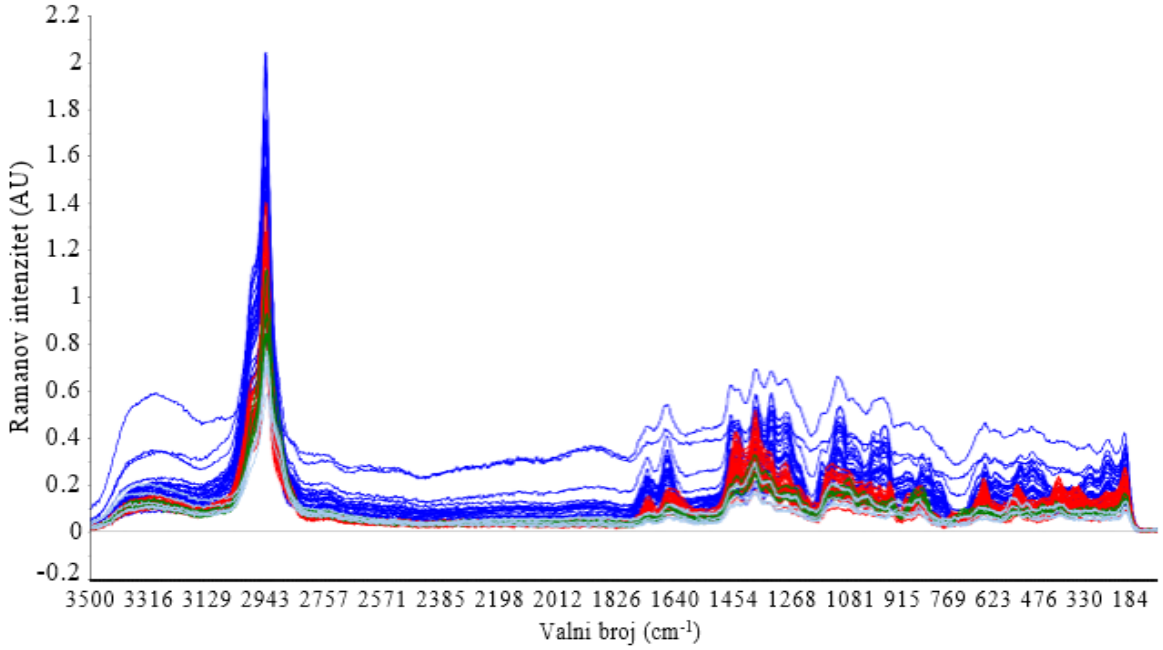
CSNS, CSPS, CEFF, TSNS, TSPS i TEFF su opisani u poglavlju 2.5.3.2.

Iz dobivenih rezultata može se zaključiti da je formirani NIR PLS-DA model izuzetno učinkovit za identifikaciju nepoznatih uzoraka PMPS A i C.

4.3. Razvoj i validacija Raman modela za identifikaciju PMPS A i PMPS C

4.3.1. Raman spektri PMPS A, C, W135 i Y

Ukupno su snimljena 73 Raman spektra PMPS A, C, W135 i Y (Slika 130.). Pri tome je snimljeno 29 spektara PMPS A, zatim 34 spektra PMPS C te po pet replikata uzoraka PMPS W135 i PMPS Y.

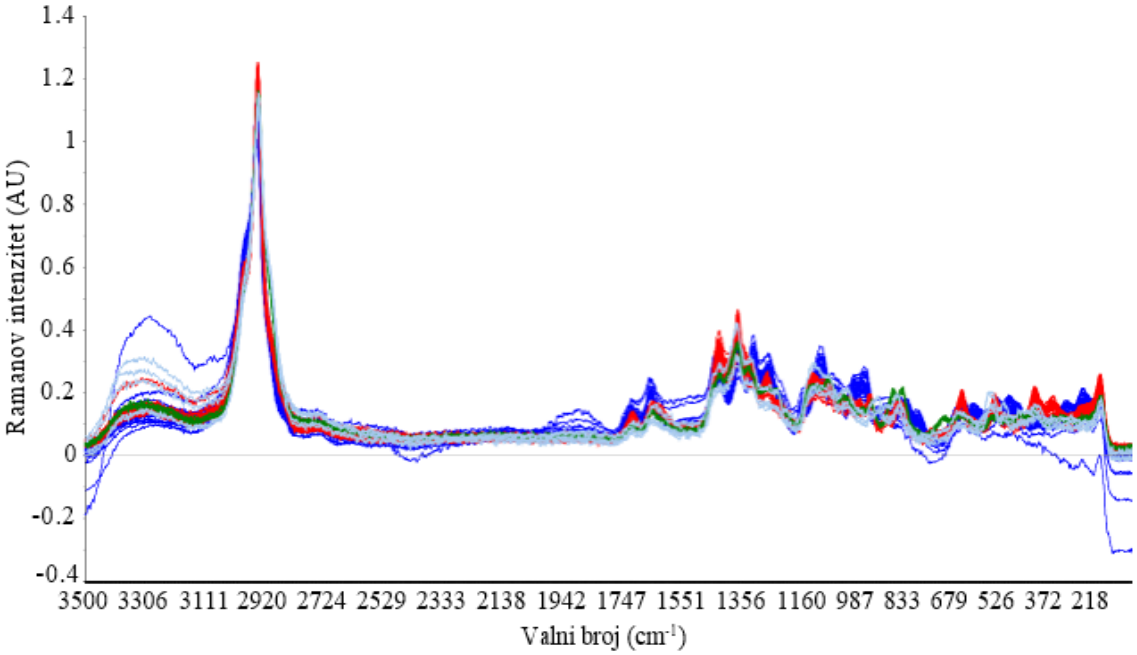


Slika 130. Raman spektri PMPS A (plavo), C (crveno), W135(zeleno) i Y(sivo) u spektralnom području $\tilde{\nu}$ = 3500 - 100 cm^{-1}.

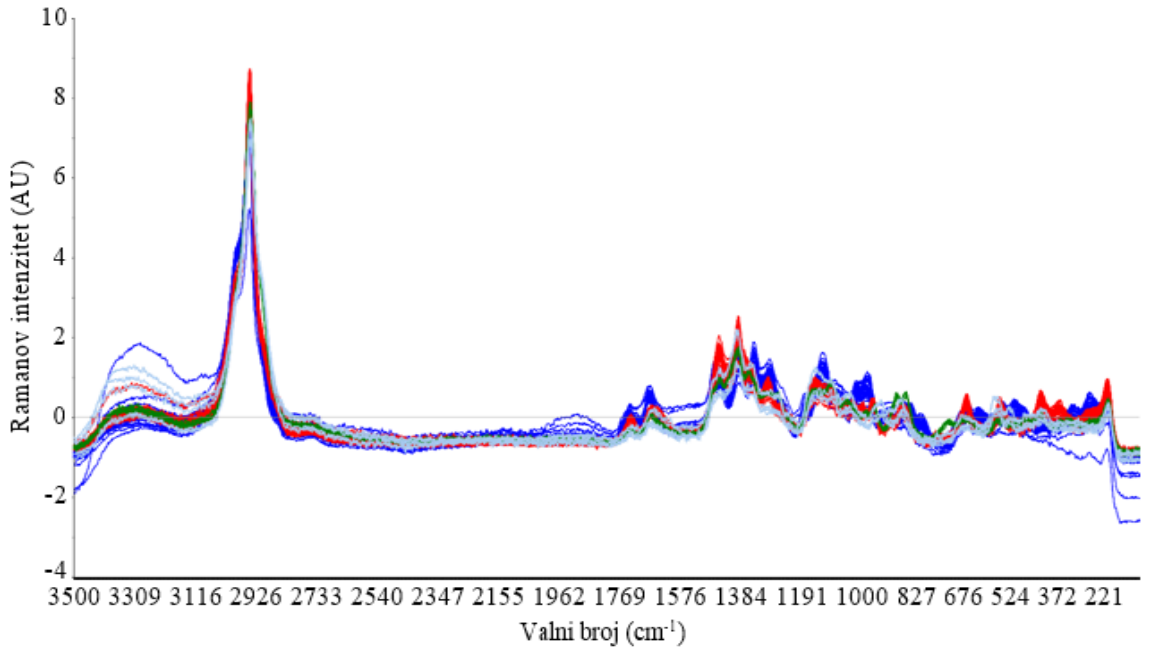
4.3.2. Kemometrijska obrada snimljenih Raman spektara PMPS A, C, W135 i Y

Različite metode matematičke predobrade Raman spektralnih podataka koristile su se kako bi se umanjile nebitne informacije i neželjene varijacije signala kao različiti šumovi te pomaci bazne linije i to sve s ciljem razvoja pouzdanog i robustnog Raman modela za identifikaciju PMPS A i C. Raman spektri PMPS A, PMPS C, PMPS W135 i PMPS Y kemometrijski su obrađeni različitim matematičkim metodama predobrade, kako je to prikazano na Slikama 131.-137. Korekcija aditivnih i multiplikativnih učinaka karakterističnih fizikalno-kemijskih

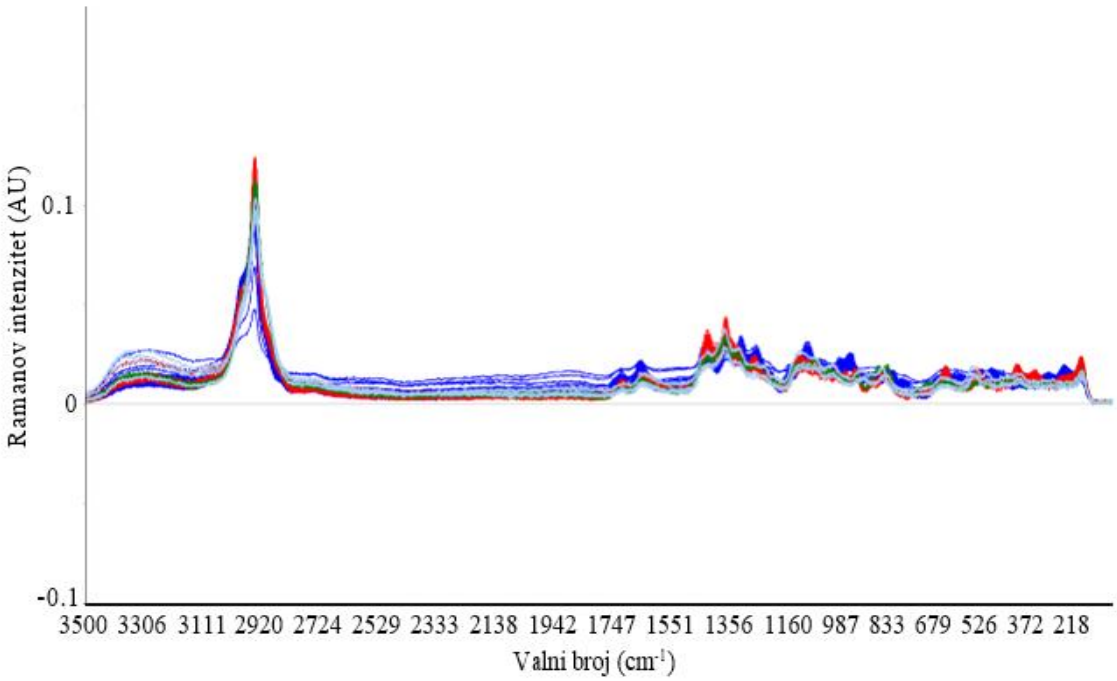
svojstava uzoraka PMPS koji proizlaze iz složenog biotehnološkog procesa proizvodnje, kao što je npr. različita veličina čestica, provedena je obradom standardnom normalnom varijatom (SNV) i višestrukom korekcijom raspršenog zračenja (MSC). Metode derivacije koristile su se za korekciju sistemskih varijacija Raman spektara i pomaka bazne linije, kao i za razdvajanje preklapajućih spektralnih vrpci. Također se za obradu Raman spektra koristila primjena Savitzky-Golay algoritma. Kako bi se svi podaci dobili na približno istoj skali, Raman spektralni podaci su se transformirali jediničnom vektorskom normalizacijom. Transformacija uklanjanja trenda primjenila se kako bi se uklonili nelinearni trendovi u Raman spektroskopskim podacima. Također, kako bi se smanjila multikolinearnost, pomak bazne linije i zakrivljenost koristila se kombinacija transformacije uklanjanja trenda sa SNV.



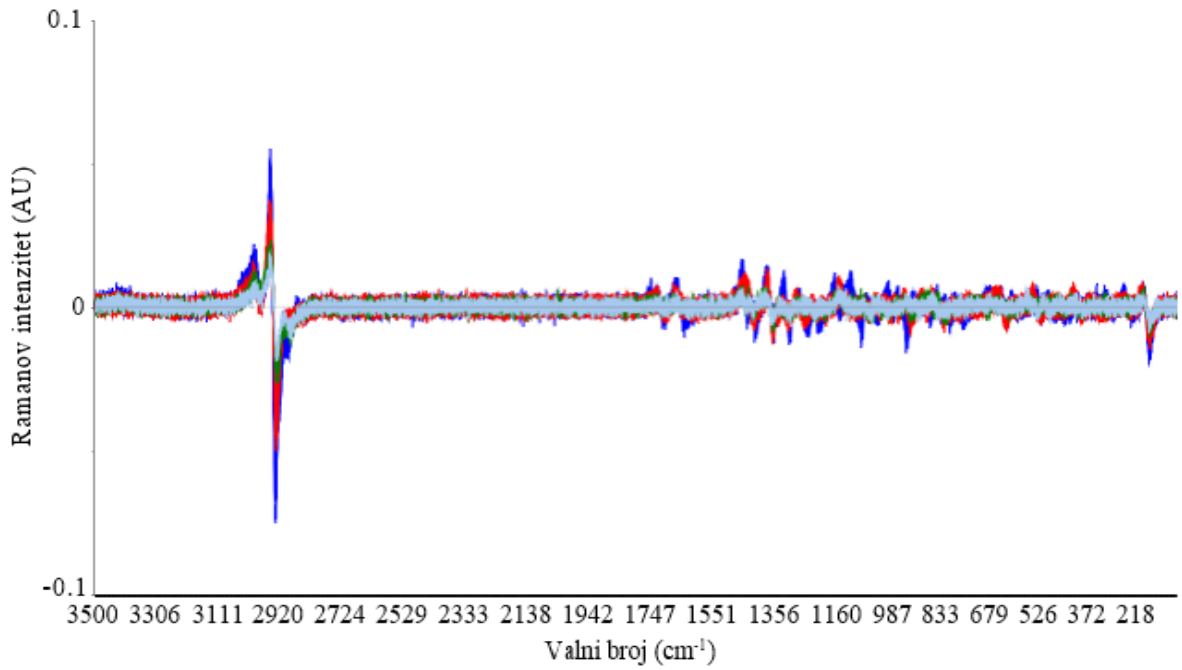
Slika 131. Raman spektri PMPS A (plavo), C (crveno), W135 (zeleno) i Y (sivo) matematički obrađeni višestrukom korekcijom raspršenog zračenja (MSC) u spektralnom području $\tilde{\nu}$ = 3500 - 100 cm^{-1}.



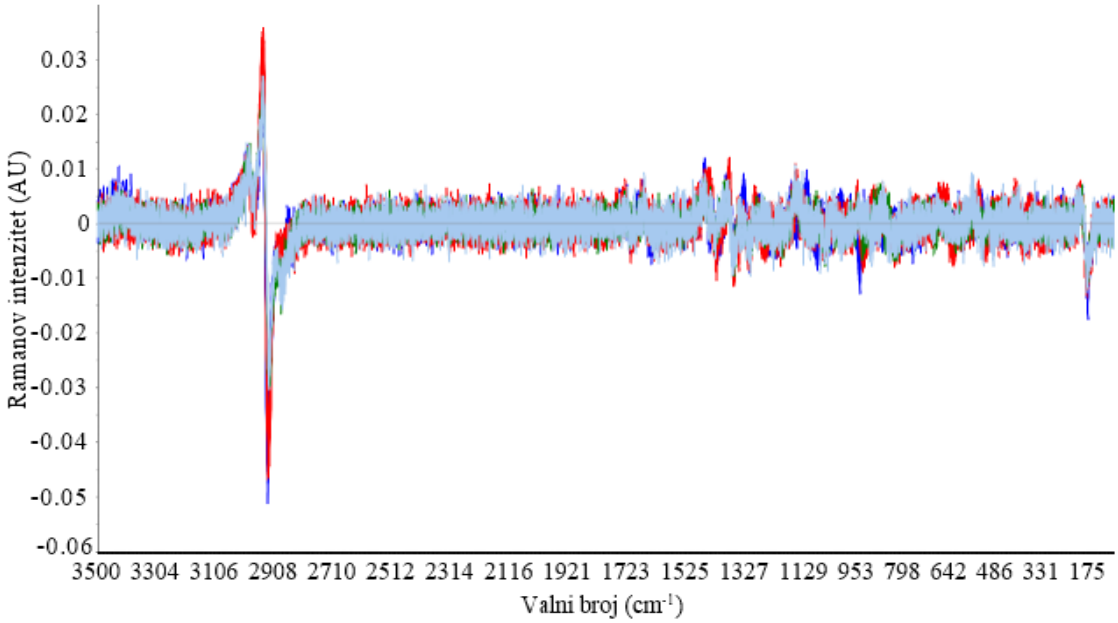
Slika 132. Raman spektri PMPS A (plavo), C (crveno), W135 (zeleno) i Y (sivo) matematički obrađeni standardnom normalnom varijatom (SNV) u spektralnom području $\tilde{\nu}$ = 3500 - 100 cm^{-1}.



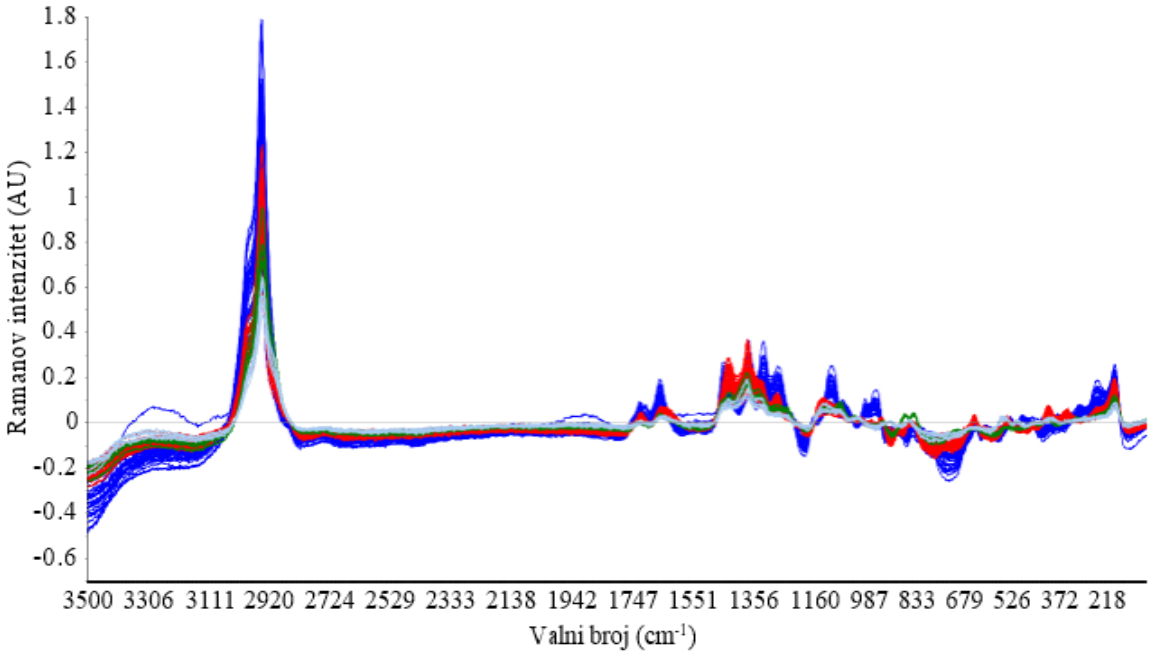
Slika 133. Raman spektri PMPS A (plavo), C (crveno), W135 (zeleno) i Y (sivo) matematički obrađeni jediničnom vektorskom normalizacijom u spektralnom području $\tilde{\nu}$ = 3500 - 100 cm^{-1}.



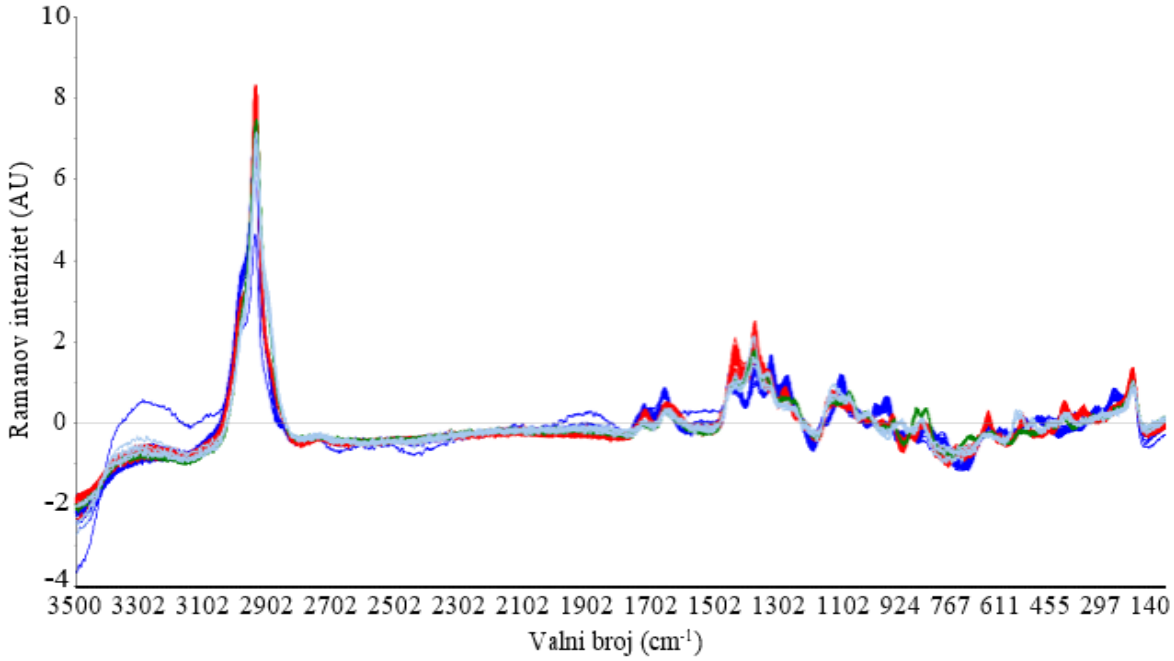
Slika 134. Raman spektri PMPS A (plavo), C (crveno), W135 (zeleno) i Y (sivo) matematički obrađeni prvom derivacijom sa Savitzky-Golay glačanjem 5.7 u spektralnom području $\tilde{\nu}$ = 3500 - 100 cm^{-1}.



Slika 135. Raman spektri PMPS A (plavo), C (crveno), W135 (zeleno) i Y (sivo) matematički obrađeni MSC-om i prvom derivacijom u spektralnom području $\tilde{\nu}$ = 3500 - 100 cm^{-1}.



Slika 136. Raman spektri PMPS A (plavo), C (crveno), W135 (zeleno) i Y (sivo) matematički obrađeni uklanjanjem trenda polinomom 4. stupnja u spektralnom području $\tilde{\nu}$ = 3500 - 100 cm^{-1}.



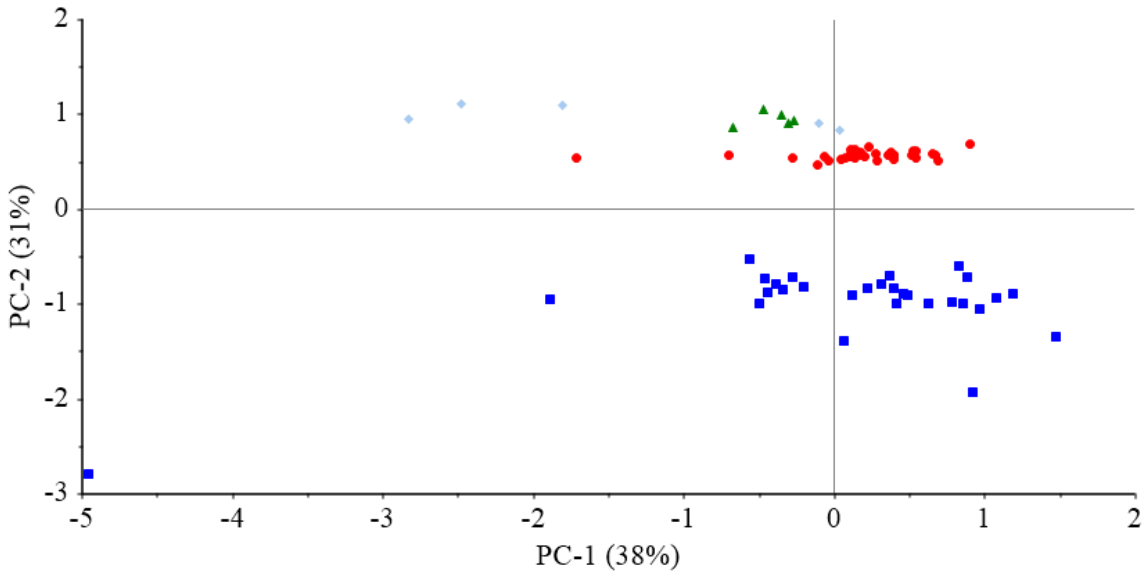
Slika 137. Raman spektri PMPS A (plavo), C (crveno), W135 (zeleno) i Y (sivo) matematički obrađeni SNV i uklanjanjem trenda polinomom 4. stupnja u spektralnom području $\tilde{\nu}$ = 3500 - 100 cm^{-1}.

Slika 137 prikazuje da kod ovako obrađenih Raman spektara PMPS A, C, W135 i Y (SNV i uklanjanje trenda polinomom 4. stupnja) se jasno vidi međusobno razdvajanje ovih četiriju polisaharida, kao što je kasnije pokazano i kod analize glavnih komponenti (PCA, Slika 144.).

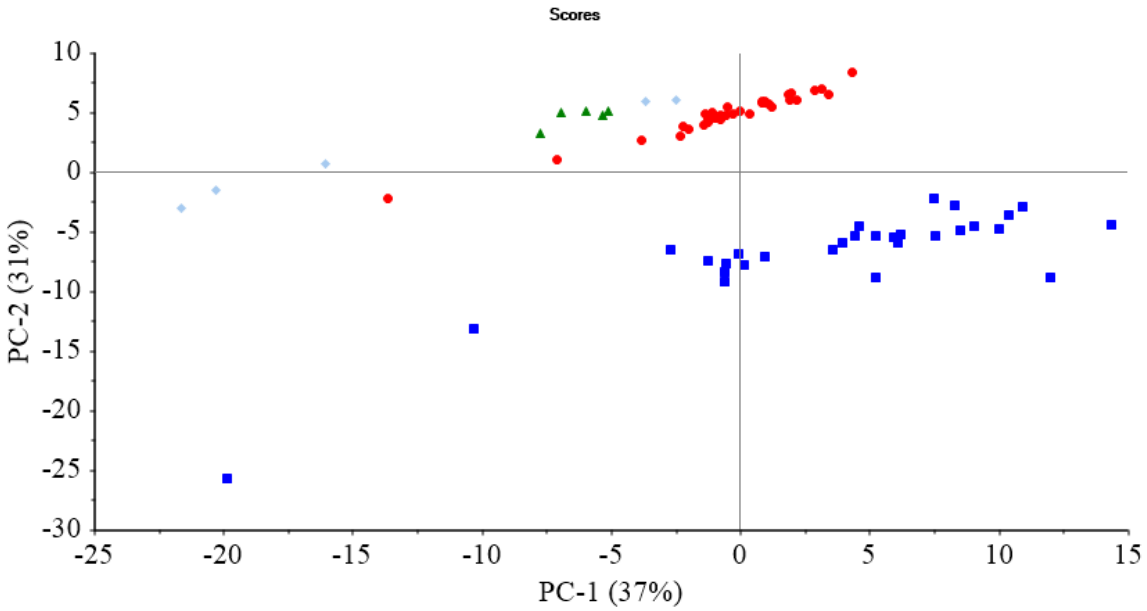
4.3.3. Eksploracijska analiza Raman spektralnih podataka PMPS A, C, W135 i Y

4.3.3.1. PCA

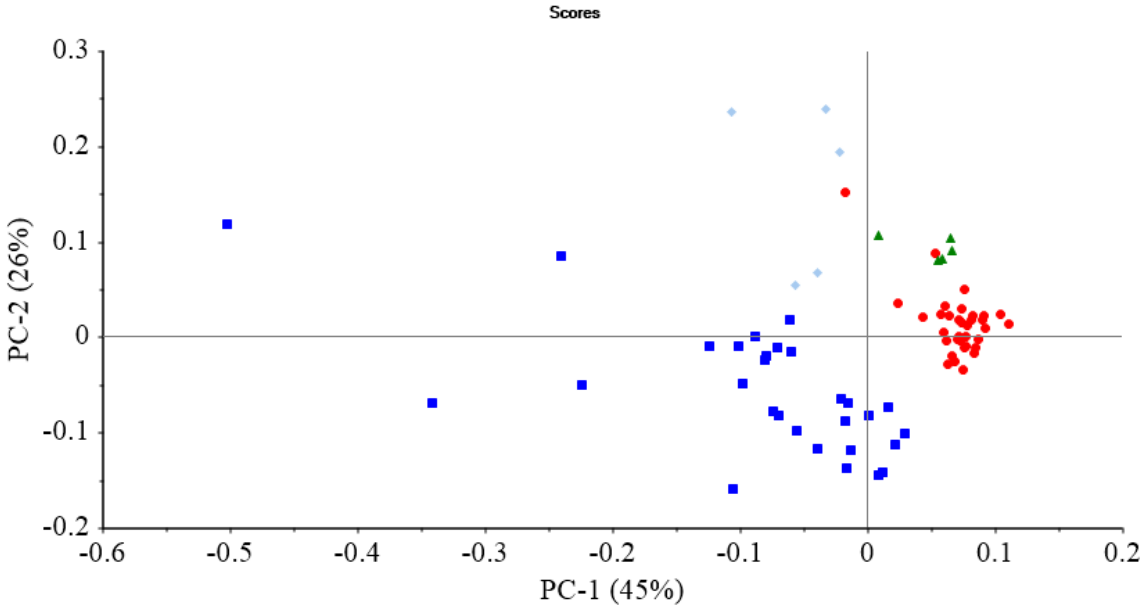
Eksploracijska analiza glavnih komponenti (PCA) Raman spektralnih podataka PMPS A, C, W135 i Y napravljena je uz korištenje Raman spektara obrađenih različitim matematičkim metodama (SNV, MSC, jediničnom vektorskom normalizacijom, prvom derivacijom, Savitzky-Golay glačanjem, uklanjanjem trenda polinomom 4. stupnja). Slično kao kod obrade NIR spektralnih podataka, PCA Raman spektralnih podataka je provedena u svrhu procjene strukture Raman podataka u multidimenzionalnom prostoru, te vizualiziranja trendova i identificiranja karakterističnih grupa u Raman podacima. Dobiveni statistički podaci vizualizirani su pomoću glavnih komponenti (PC, vidi poglavlje 2.5.2.1. u Općem dijelu), kako je to pokazano na Slikama 138. - 145.



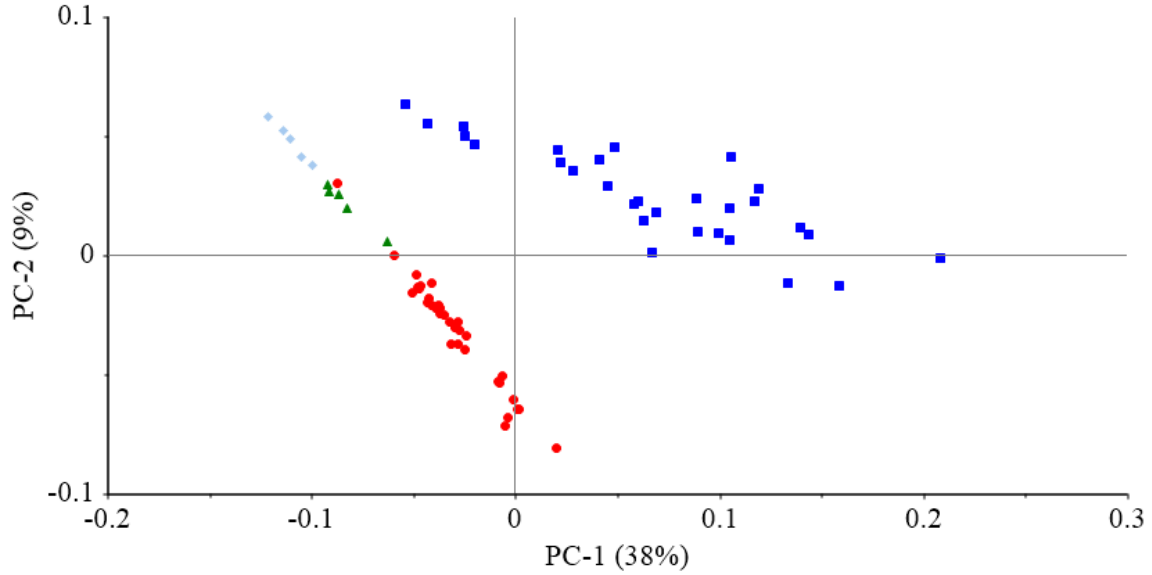
Slika 138. Raspodjela faktorskih bodova PC1 i PC2 matematički obrađenih (MSC) Raman spektara PMPS A, C, W135 i Y. Faktorski bodovi PC1 i PC2 se odnose na PMPS: A (plavi kvadrati), C (crveni krugovi), W135 (zeleni trokuti) i Y (sivi rombovi).



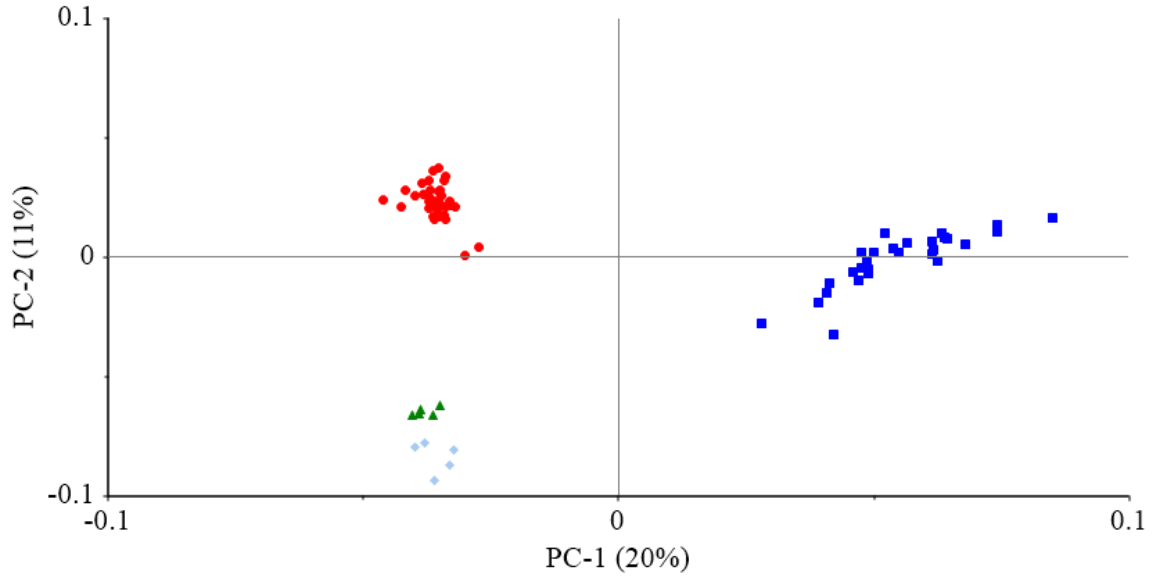
Slika 139. Raspodjela faktorskih bodova PC1 i PC2 matematički obrađenih (SNV-om) Raman spektara PMPS A, C, W135 i Y. Faktorski bodovi PC1 i PC2 se odnose na PMPS: A (plavi kvadrati), C (crveni krugovi), W135 (zeleni trokuti) i Y (sivi rombovi).



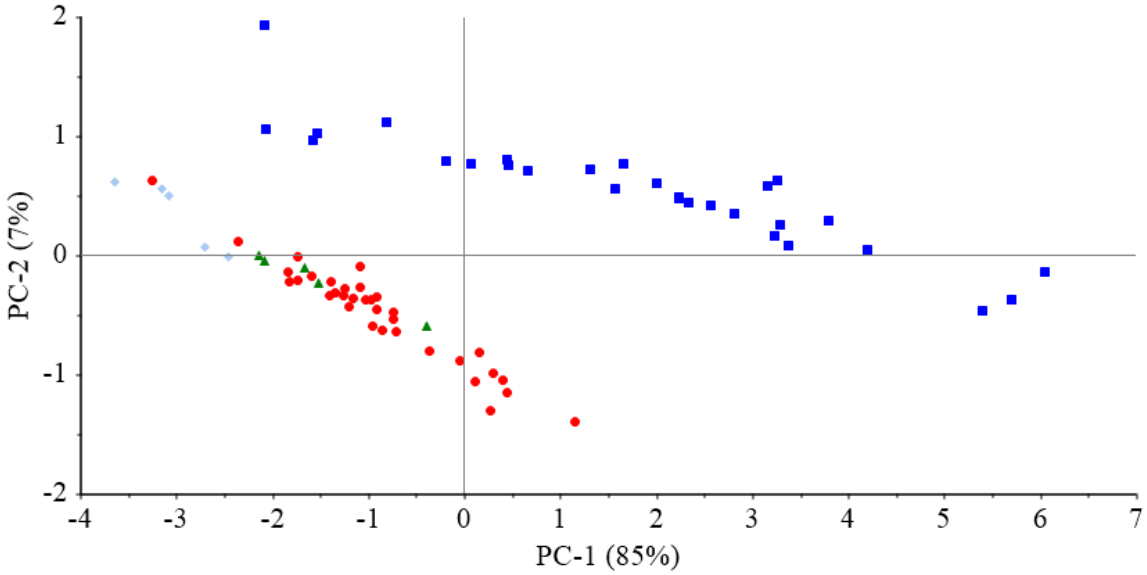
Slika 140. Raspodjela faktorskih bodova PC1 i PC2 matematički obrađenih (jediničnom vektorskom normalizacijom) Raman spektara PMPS A, C, W135 i Y. Faktorski bodovi PC1 i PC2 se odnose na PMPS: A (plavi kvadrati), C (crveni krugovi), W135 (zeleni trokuti) i Y (sivi rombovi).



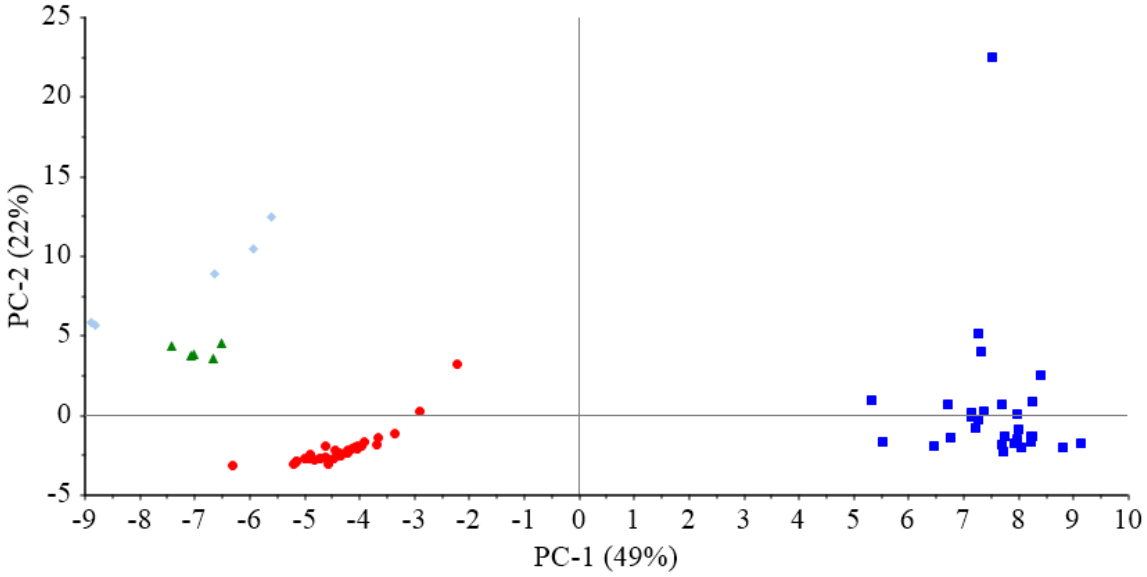
Slika 141. Raspodjela faktorskih bodova PC1 i PC2 matematički obrađenih (prva derivacija i Savitzky-Golay glačanje 5.7) Raman spektara PMPS A, C, W135 i Y. Faktorski bodovi PC1 i PC2 se odnose na PMPS: A (plavi kvadrati), C (crveni krugovi), W135 (zeleni trokuti) i Y (sivi rombovi).



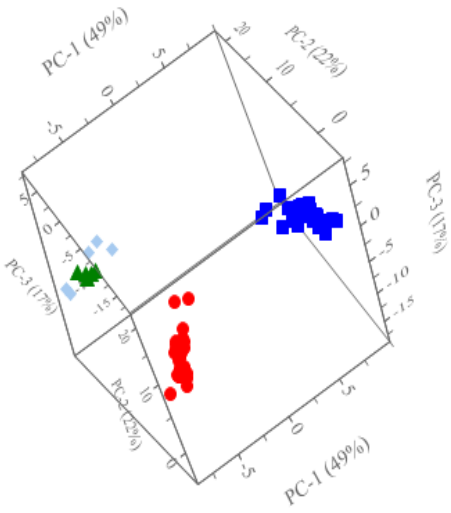
Slika 142. Raspodjela faktorskih bodova PC1 i PC2 matematički obrađenih (MSC-om i prvom derivacijom) Raman spektara PMPS A, C, W135 i Y. Faktorski bodovi PC1 i PC2 se odnose na PMPS: A (plavi kvadrati), C (crveni krugovi), W135 (zeleni trokuti) i Y (sivi rombovi).



Slika 143. Raspodjela faktorskih bodova PC1 i PC2 matematički obrađenih (uklanjanje trenda polinomom 4. stupnja) Raman spektara PMPS A, C, W135 i Y. Faktorski bodovi PC1 i PC2 se odnose na PMPS: A (plavi kvadrati), C (crveni krugovi), W135 (zeleni trokuti) i Y (sivi rombovi).



Slika 144. Raspodjela faktorskih bodova PC1 i PC2 matematički obrađenih (SNV i uklanjanje trenda polinomom 4. stupnja) Raman spektara PMPS A, C, W135 i Y. Faktorski bodovi PC1 i PC2 se odnose na PMPS: A (plavi kvadrati), C (crveni krugovi), W135 (zeleni trokuti) i Y (sivi rombovi).



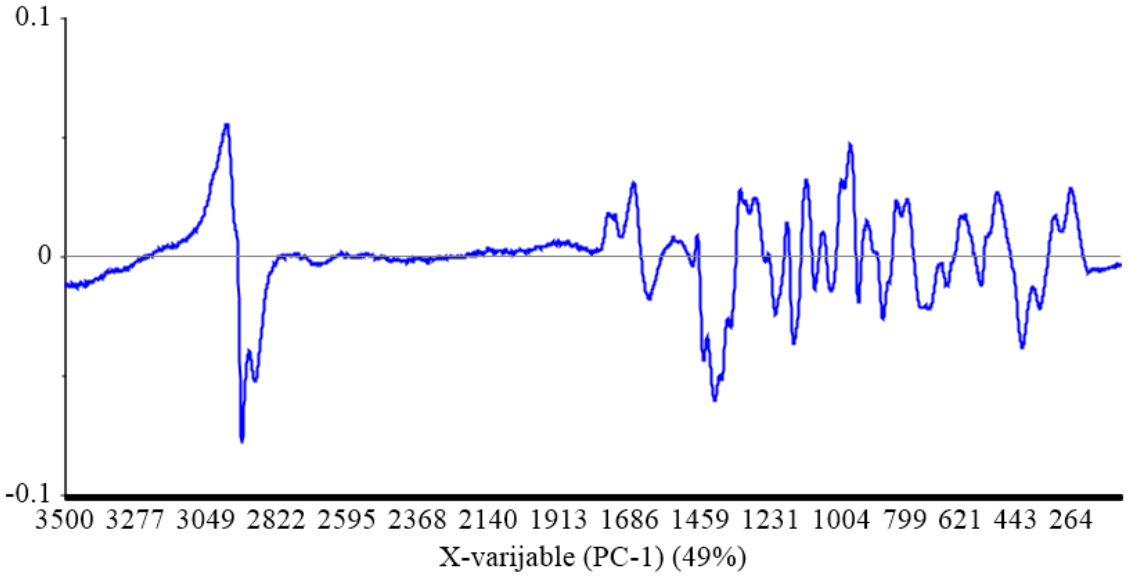
Slika 145. Trodimenzionalni prikaz faktorskih bodova PC1, PC2 i PC3 matematički obrađenih (SNV-om i uklanjanje trenda polinomom 4. stupnja) Raman spektara PMPS A, C, W135 i Y. Faktorski bodovi PC1 i PC2 i PC3 se odnose na PMPS: A (plavi kvadrati), C (crveni krugovi), W135 (zeleni trokuti) i Y (sivi rombovi).

Rezultati PCA anlize nakon provedenih različitih matematičkih predobrada Raman spektralnih podataka prikazani su na odgovarajućim slikama faktorskih bodova (PC, Slike 138. - 145.). Kod prve dvije glavne komponente - PC1 i PC2 (Slike 138. - 144.) jasno se vidi grupiranje i razdvajanje pročišćenih meningokoknih polisaharida različitih serogrupa nakon prethodne obrade kombinacijom SNV-a i ulanjanjem trenda polinomom 4. stupnja. Na prikazu faktorskih bodova u dvodimenzionalnom prostoru (Slika 144.) može se vidjeti razdvajanje PMPS A od PMPS C te da su PMPS W135 i Y grupirani zajedno, ali jasno razdvojeni od PMPS A i C i to nakon što su snimljeni Raman spektri prethodno obrađeni kombinacijom SNV i uklanjanjem trenda polinomom 4. stupnja. Matematička obrada Raman spektara SNV-om je korištena kako bi se učinkovito uklonile multiplikativne interferencije raspršenja i veličine čestica PMPS, dok se uklanjanje trenda koristlo u svrhu korekcije bazne linije i nelinearnosti Raman spektara. Tako matematički obrađeni Raman spektri PMPS korišteni su u daljnjem formiranju dvaju Raman modela - SIMCA i PLS-DA.

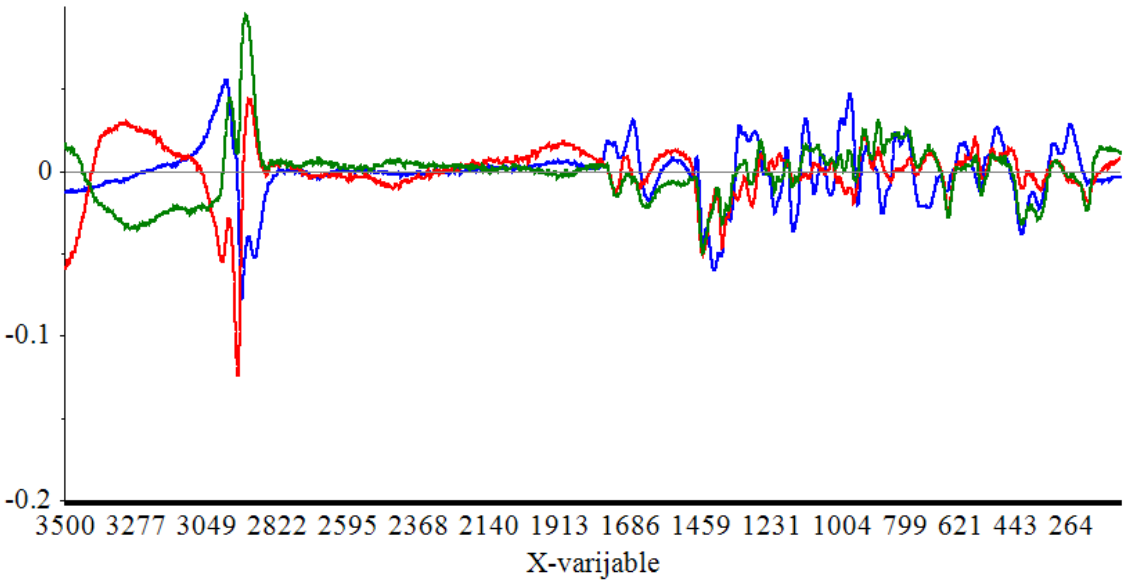
Na prikazu faktorskih bodova u trodimenzionalnom prostoru (Slika 145.), prve tri glavne komponente (PC1, PC2 i PC3) uključuju 49,0 %, 22,0 % i 17,0 % varijance Raman spektralnih

podataka. Ovi su podaci dobiveni iz ukupno 73 Raman spektra PMPS A, C, W135 i Y, i zajedno opisuju 88,0 % varijance ovih Raman spektralnih podataka, koja se također, kao i kod NIR spektralnih podataka, odnosi na kemijski sastav uzoraka različitih PMPS.

Kako bi se definirale najvažnije Raman spektralne regije, koje su ključne za međusobno razlikovanje ovih četiriju polisaharida (PMPS A, C W135 i Y), napravljen je profil opterećenja. (Slike 146. i 147.).



Slika 146. PC1 opterećenja po valnim brojevima ($\tilde{\nu}$) dobivena PCA analizom matematički obrađenih (SNV i uklanjanje trenda polinomom 4. stupnja) Raman spektara PMPS A, C, W135 i Y.

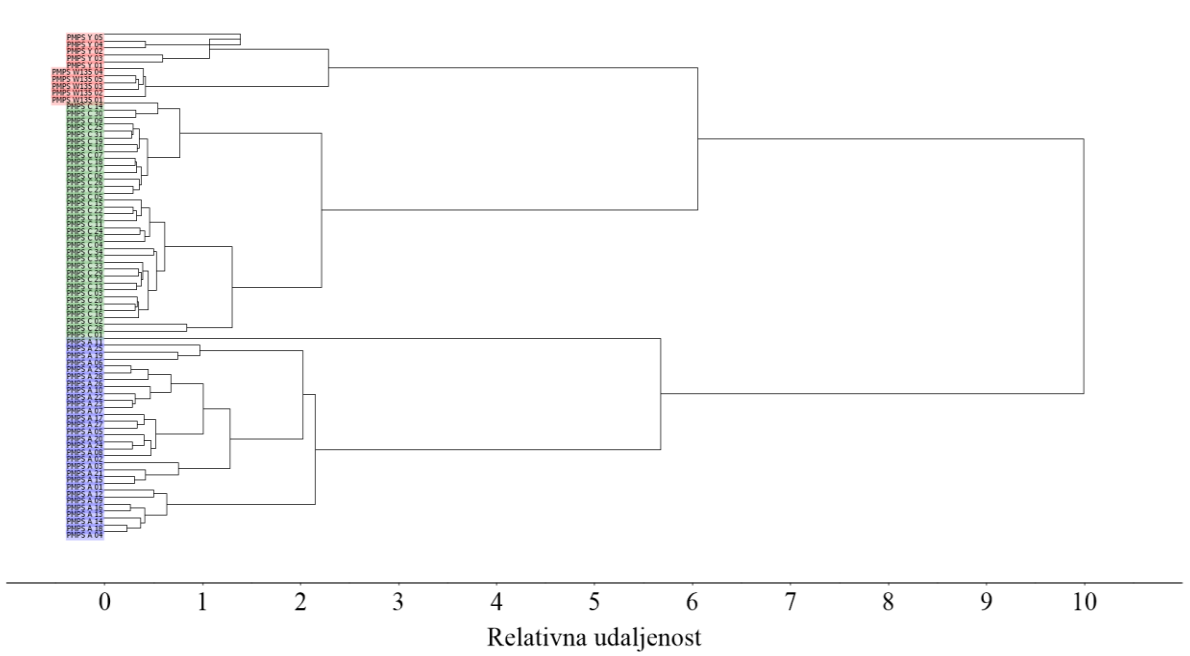


Slika 147. Profil opterećenja za PC 1 (plava), PC 2 (crvena) i PC 3 (zelena) matematički obrađenih (SNV i uklanjanje trenda polinomom 4. stupnja) Raman spektara PMPS A, C, W135 i Y.

Izradom profila opterećenja za PC1 identificirane su najkarakterističnije spektralne regije u Raman spektrima PMPS A, C, W135 i Y, odgovorne za međusobno razlikovanje serogrupa pročišćenih meningokoknih polisaharida. Karakteristična spektralna regija u rasponu valnih brojeva oko $\tilde{\nu}$ = 2980 cm^{-1} koje proizlaze od C-H i CH_2 istezanja; $\tilde{\nu}$ = 2930 cm^{-1} proizlaze od CH_2 asimetričnog istezanja; $\tilde{\nu}$ = 2875 cm^{-1} proizlaze od CH_2 simetričnog istezanja ; $\tilde{\nu}$ = 1750 - 1700 cm^{-1} C=O vibracija istezanja karbonilne grupe te $\tilde{\nu}$ = 1680 - 1639 cm^{-1} proizlaze od C=ONHR istezanja; $\tilde{\nu}$ = 1500 - 1200 cm^{-1} C-H/CH_2 deformacijske vibracije; $\tilde{\nu}$ = 1350 - 1140 P=O istezanja; $\tilde{\nu}$ = 1200 - 950 cm^{-1} proizlaze od C-C i C-O simetričnog istezanja; $\tilde{\nu}$ = 1150 - 1070 cm^{-1} proizlazi od C-O-C asimetričng istezanja; $\tilde{\nu}$ = 980 - 965 cm^{-1} proizlaze od C-H savijanja izvan ravnine; $\tilde{\nu}$ = 800 - 100 deformacijske vibracije C-C-O grupa (Larkin, 2018).

Nadalje, metodom hijerarhijskog klasteriranja (Poglavlje 4.3.3.2.), grupirani su slični Raman spektri PMPS A, C, W135 i Y (ovdje ispod).

4.3.3.2. Klasterska analiza



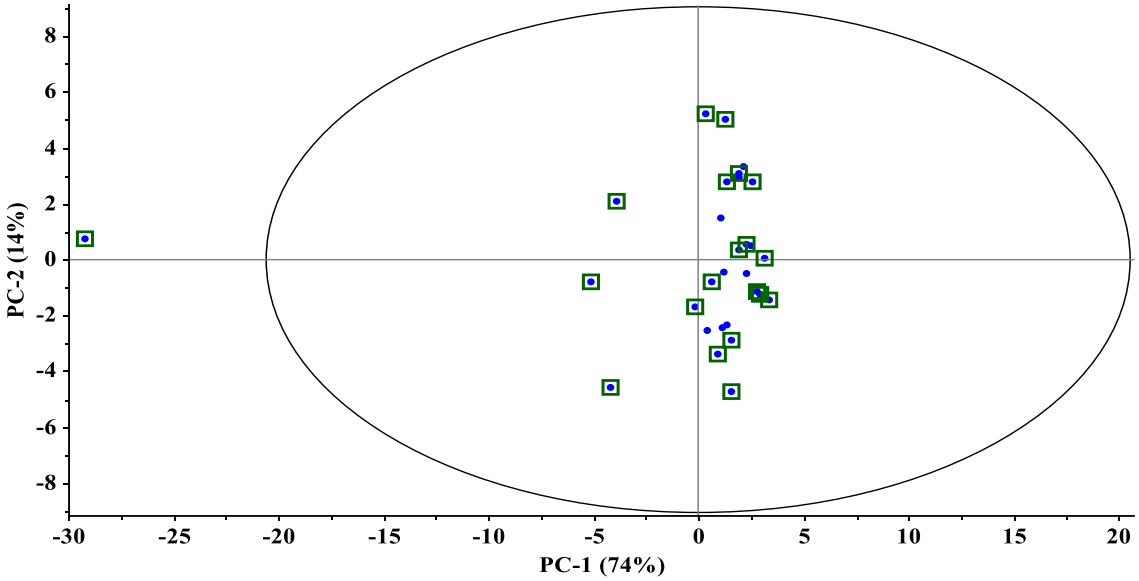
Slika 148. Klasterska analiza uzoraka PMPS A, C, W135 i Y dobivena Ward algoritmom za hijerarhijsko grupiranje Raman spektralnih podataka u spektralnom području $\tilde{\nu}$ = 3500 - 100 cm^{-1}. Analizom dendograma mogu se identificirati tri klastera uzoraka PMPS: klaster 1 (plavo) grupirao je uzorke PMPS A, klaster 2 (zeleno) uzorke PMPS C, dok je klaster 3 (crveno) grupirao uzorke W135 i Y.

Iz ovih se rezultata može zaključiti da se klasifikacija uzoraka PMPS temelji na njihovoj karakterističnoj kemijskoj strukturi. Grupiranjem uzoraka Ward klasterirajućom metodom (Slika 148.) potvrđena je PCA eksploracijska analiza, kako je to prikazano prethodno na Slikama 144. - 145.

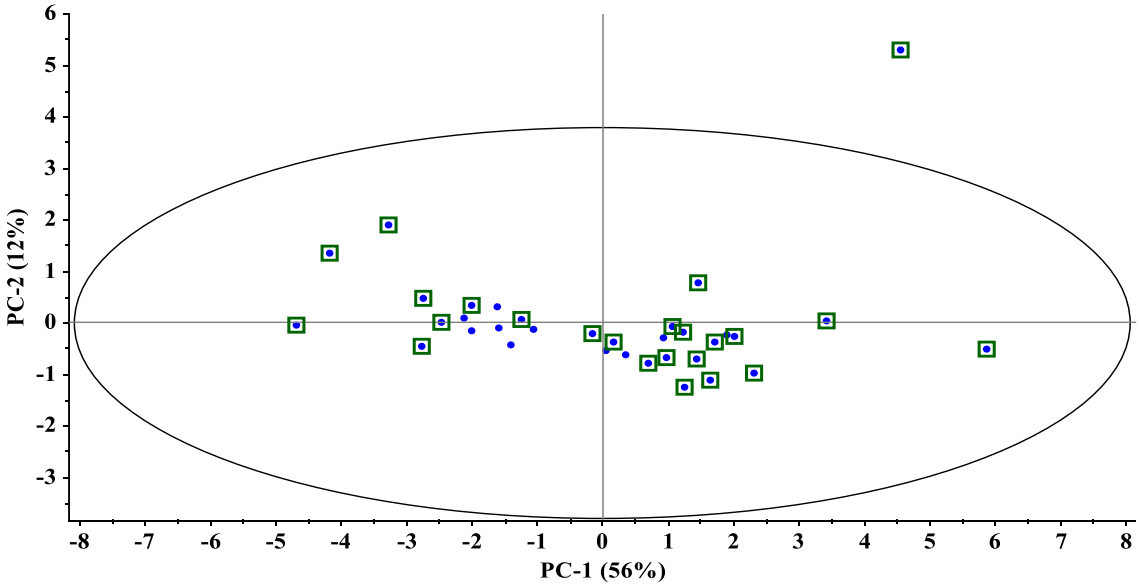
4.3.4. Raspodjela Raman spektara PMPS A, C, W135 i Y u kalibracijski i evaluacijski set

Kako bi se procijenila prediktivna sposobnost Raman modela u cilju identifikacije nepoznatih uzoraka PMPS A i PMPS C, provedena je validacija formiranih Raman (SIMCA i PLS-DA) modela. Postupak validacije ovih Raman modela zahtijeva podjelu uzoraka PMPS na setove, koji će se u tu svrhu koristiti. Uzorci PMPS A i C podjeljeni su na (1) trening (kalibracijski) set, koji se koristio za kalibraciju i optimizaciju formiranih Raman modela i to postupkom

unakrsne validacije, i (2) vanjski test set, koji će se koristiti za validaciju optimiranih Raman modela. Kako bi se izbjegla precijenjena sposobnost predviđanja Raman modela, uzorci vanjskog test seta nisu sudjelovali u formiranju i optimizaciji Raman modela. Na Slikama 91. i 92. prikazana je podjela uzoraka PMPS A i C na trening (kalibracijski) set i test set. Uzorci su razdijeljeni na temelju analize njihove udaljenosti u trodimenzionalnom PCA prostoru prema uniformnom Kennard-Stone algoritmu (Slike 149. i 150.).



Slika 149. Raspodjela uzoraka PMPS A Kennard-Stone algoritmom na trening i test set.Trening set (zeleni kvadrati), test set (plave točke).



Slika 150. Raspodjela uzoraka PMPS C Kennard-Stone algoritmom na trening i test set. Trening set (zeleni kvadrati), test set (plave točke).

Na Slikama 149. i 150 može se vidjeti jednolika raspodjela uzoraka trening i test seta te je tako obuhvaćena ukupna varijabilnost uzoraka oba seta, kako za PMPS A tako i za PMPS C.

4.3.4.1. Raspodjela Raman spektara PMPS A u kalibracijski i evaluacijski set

Ukupno 29 Raman spektra od 20 serija PMPS A podijeljeni su na (1) kalibracijski (trening) set, koji je sadržavao ukupno 20 snimljenih Raman spektara od ukupno 13 serija, i (2) evaluacijski (test) set, koji je sadržavao devet Raman spektara snimljenih kod sedam proizvodnih serija PMPS A.

4.3.4.2. Raspodjela Raman spektara PMPS C u kalibracijski i evaluacijski set

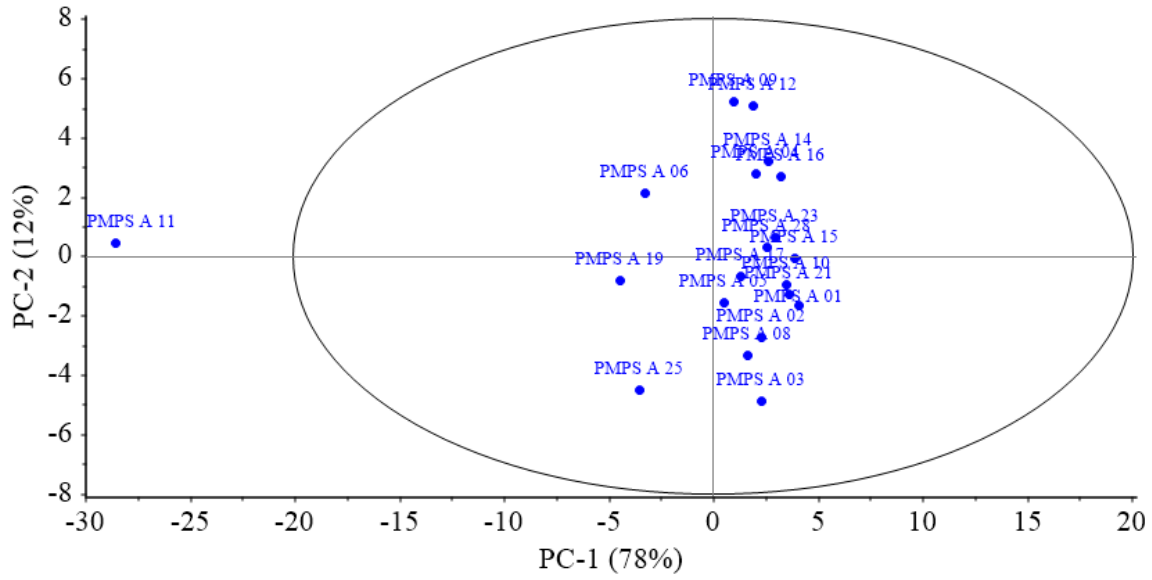
Ukupno 34 Raman spektra od 23 serije uzoraka PMPS C podijeljeni su u (1) kalibracijski (trening) set, koji se sastojao od 24 spektra (ukupno 16 serija PMPS C), i (2) evaluacijski (test) set, koji se sastoji od ukupno 10 Raman spektara od preostalih sedam serija PMPS C.
Pet replikata uzoraka PMPS W135 i Y korišteni su kao negativna proba za PMPS A i C.

4.3.5. Raman SIMCA model

Kao prvi korak u formiranju Raman SIMCA modela, provedene su pojedinačne analize glavnih komponenti (PCA), posebno za PMPS A, a posebno za PMPS C.

4.3.5.1. PCA modeliranje Raman spektara PMPS A

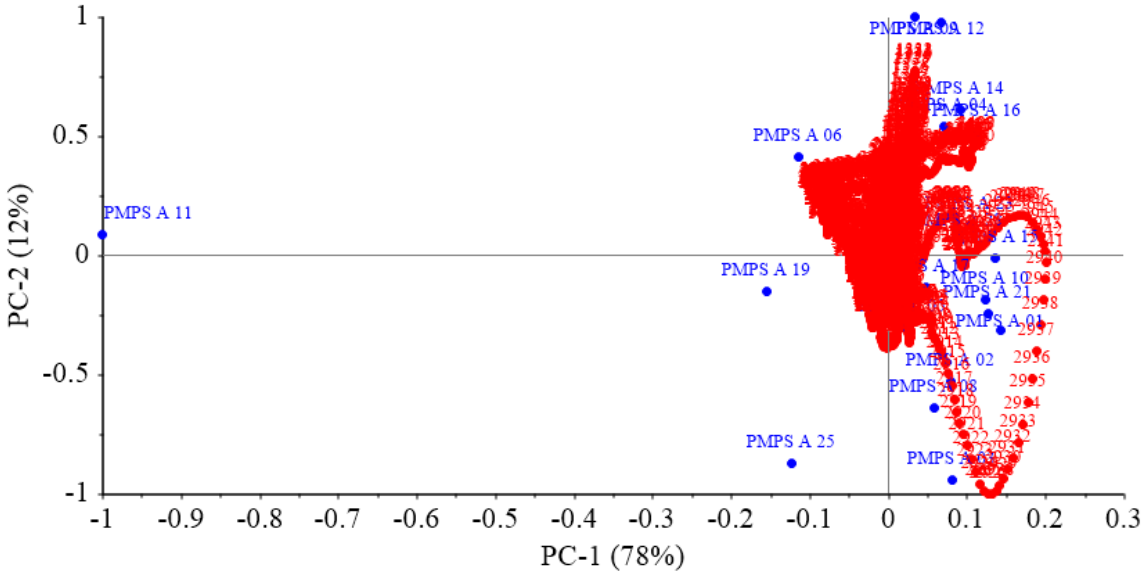
PCA je provedena uz korištenje kalibracijskog seta PMPS A, koji je buhvatio 20 Raman spektara uzoraka ovog polisaharida. Tako je formiran PCA model, a pomoću istog seta uzoraka provela se i optimizacija PCA modela postupkom unakrsne validacija i određen je optimalan broj PC faktora.
Ponajprije je načinjena raspodjela faktorskih bodova (Slika 151.).



Slika 151. Raspodjela faktorskih bodova PC1 i PC2 matematički obrađenih (SNV i uklanjanje trenda polinomom 4. stupnja) Raman spektara PMPS A u spektralnom području $\tilde{\nu}$ = 3500 - 100 cm^{-1}. Hotelling T^2 elipsa označava 95 %-tni interval pouzanosti.

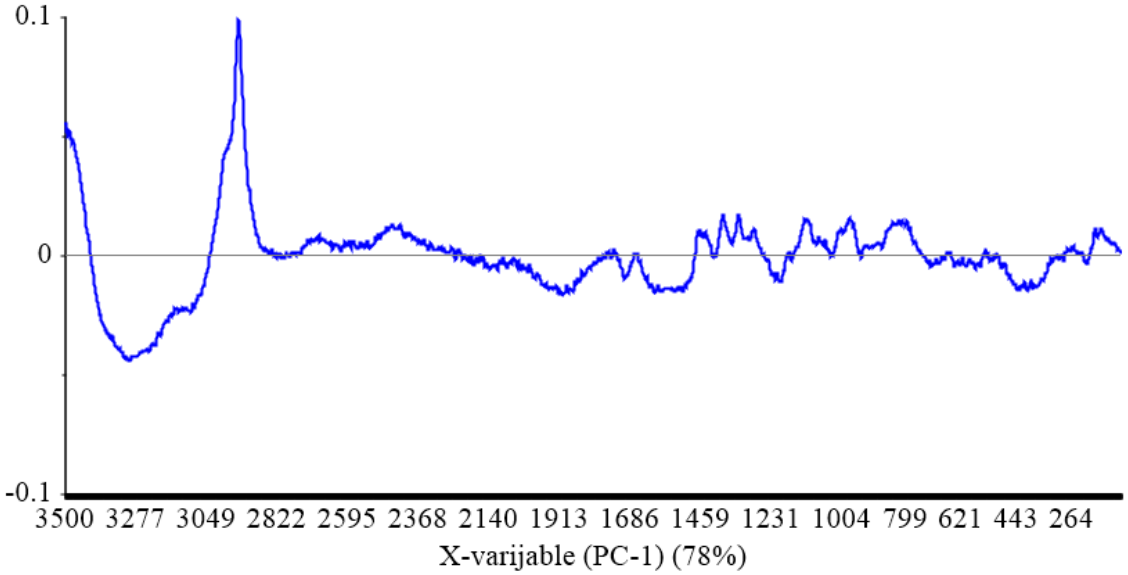
Na Slici 151. prikazani su faktorski bodovi prve i druge glavne komponente (PC1 i PC2) za uzorke PMPS A. Ovdje se jasno može vidjeti struktura podataka i njihove međusobne korelacije. Faktorski bodovi koji su blizu jedan drugog ukazuju na slične uzorke, odnosno pozitivno koreliraju, a dijametralno suprotni uzorci imaju negativnu korelaciju. Na Slici 151. se može vidjeti jedan ekstremni uzorak tj. uzorak PMPS A 11, koji je izvan Hotelling T^2 elipse (interval pouzdanosti 95 %). Ovaj je uzorak bilo potrebno dalje statistički analizirati kao eventualni netipični uzorak. Svi ostali uzorci PMPS A su jednoliko raspodijeljeni unutar područja Hotelling T^2 elipse. Prva glavna komponenta (PC1) obuhvaća 78 % varijance, a druga glavna komponenta (PC2) 12 % varijance. Odnosi među varijablama prikazani su faktorskim opterećenjima (Slike 152 .- 155.).

Analiza opterećenja vrlo je slična analizi faktorskih bodova. Svaka varijabla ima opterećenje na svakom PC-u. Prikaz opterećenja ukazuje na to koliko varijabla utječe na PC, odnosno koliko varijabla pridonosi tom PC-u.

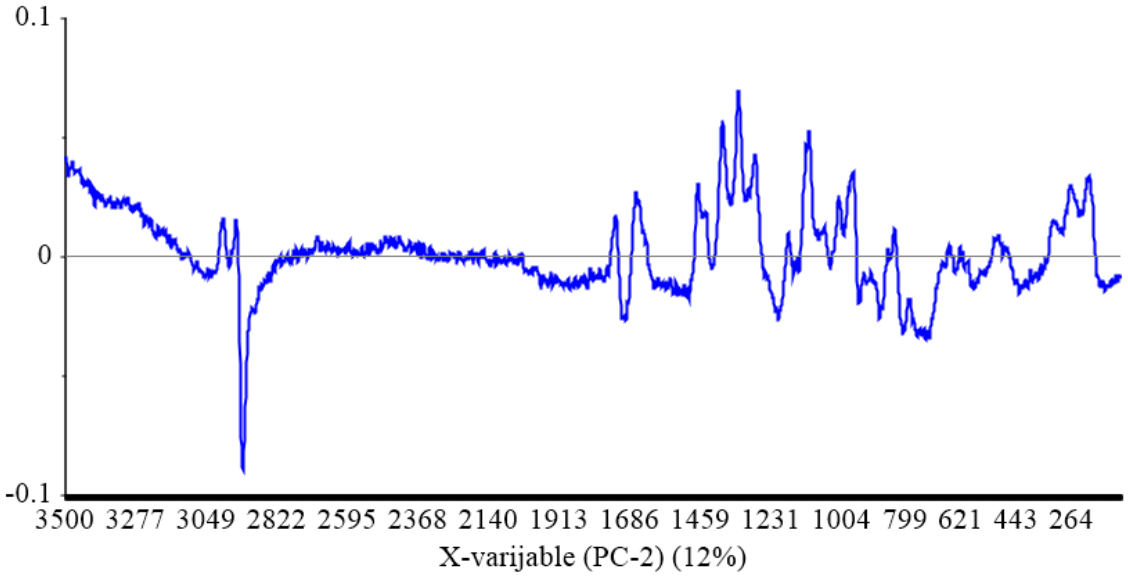


Slika 152. Raspodjela faktorskih bodova PC1 i PC2 matematički obrađenih (SNV i uklanjanje trenda polinomom 4. stupnja) Raman spektara PMPS A u spektralnom području $\tilde{\nu}$ = 3500 - 100 cm^{-1} sa PC1 i PC2 opterećenjem. Opterećenja su označena crvenom bojom, a uzorci plavom bojom.

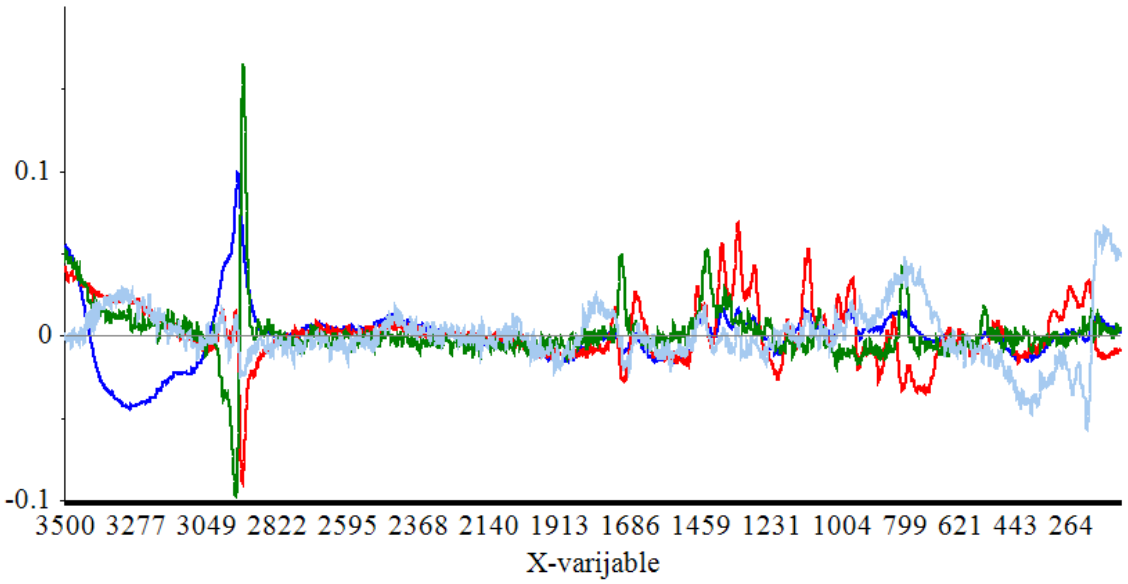
Ukoliko se promatraju uzorci PMPS A i pripadajuće varijable u međusobnom odnosu (Slika 152.), moguće je interpretirati svojstva ovih uzorka. Naime, pomoću linijskih opterećenja, koja su najprikladnija za interpretaciju varijabli Raman spektralnih podataka radi sličnosti prikaza opterećenja sa izvornim spektralnim podacima, utvrđene su varijable, odnosno spektralne regije koje imaju najveći utjecaj na pojedinu glavnu komponentu.



Slika 153. PC1 opterećenja po valnim brojevima ($\tilde{\nu}$) dobivena PCA analizom Raman spektara PMPS A.



Slika 154. PC2 opterećenja za po valnim brojevima ($\tilde{\nu}$) dobivena PCA analizom Raman spektara PMPS A.

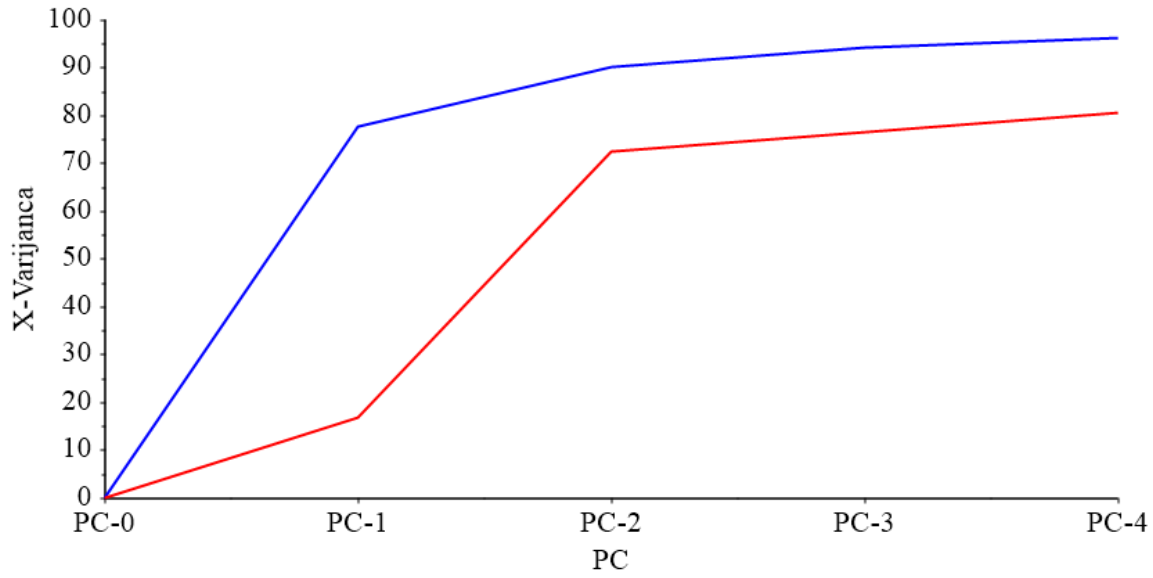


Slika 155. PC1 (plavo), PC2 (crveno), PC3 (zeleno) i PC4 (sivo) opterećenja po valnim brojevima ($\tilde{\nu}$).

Preklopljena linijska opterećenja za PC1, PC2 i PC3 jasno prikazuju Raman spektralne regije s najviše utjecaja na pojedine glavne komponente. (Slika 155.), odnosno linijska opterećenja na Slikama 153.-155. prikazuju najutjecajnije varijable u formiranju PCA modela.

Na slikama 153-155. mogu se vidjeti najutjecajnije varijable u formiranju PMPS A modela u regijama oko $\tilde{\nu}$ = 2940 cm^{-1} proizlaze od CH_2 asimetričnog istezanja; $\tilde{\nu}$ = 1750 - 1630 cm^{-1} proizlaze od C=ONHR vibracija istezanja; $\tilde{\nu}$ = 1500 - 1200 cm^{-1} C-H/CH_2 deformacijske vibracije; $\tilde{\nu}$ = 1350 - 1140 cm^{-1} proizlaze od P=O istezanja; $\tilde{\nu}$ = 1200-950 cm^{-1} proizlaze od C-C i C-O simetričnog istezanja; $\tilde{\nu}$ = 1150 - 1070 cm^{-1} proizlazi od C-O-C asimetričng istezanja; $\tilde{\nu}$ = 980 - 965 cm^{-1} proizlaze od C-H savijanja van ravnine; $\tilde{\nu}$ = 800 - 100 deformacijske vibracije C-C-O grupa (Larkin, 2018). .

Pomoću kumulativne kalibracijske varijance (Slika 156.) odredila se optimalna dimenzionalnost modela, odnosno, odredio se optimalan broj PC-ova kao ključna točka u formiranju PCA modela. Od izuzetne je važnosti za identifikacijsku odnosno klasifikacijsku sposobnost modela odrediti optimalan broj PC-ova, kako ne bi došlo do uključivanja suvišnih i nepotrebnih informacija u ovaj model ili pak suprotno –nepotpune količine informacija u modelu, koje su ključne za klasifikacijsku sposobnost modela.

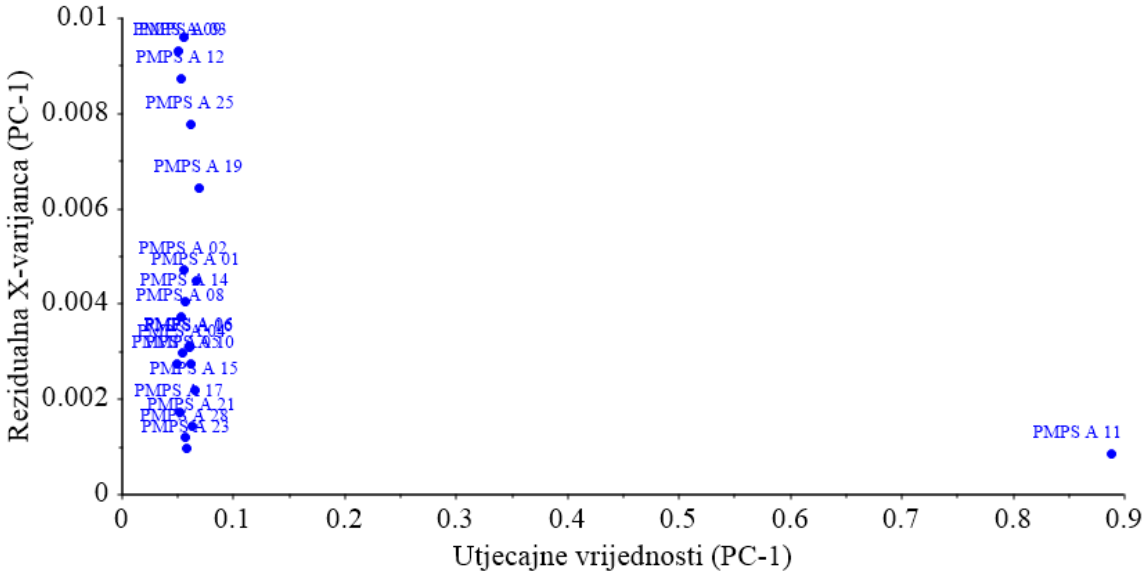


Slika 156. Kumulativna kalibracijska (plava linija) i validacijska (crvena linija) varijanca za svaki PC.

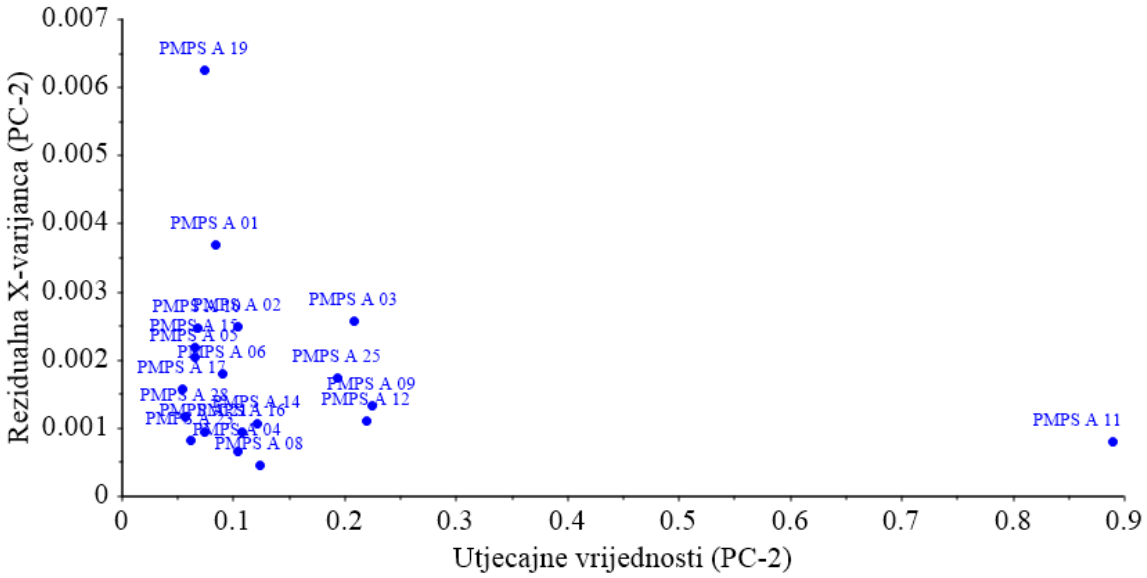
Slika 156. prikazuje koliko varijance u kalibracijskim i validacijskim podacima opisuju različite komponente. Dvije glavne kalibracijske komponente (PC) opisuju 90 % ukupne varijance (PC 1 -78 % i PC 2 -12 %) te je zbog toga formiran PCA model sa dvije glavne komponente. Slika 156. prikazuje razliku između kalibracijske i validacijske kumulativne varijance. Može se pretpostaviti prisutnost netipičnih ili ekstremnih uzoraka u kalibracijskom skupu uzoraka PMPS A. Relativno je mali broj podataka ovdje korišten za PMPS A i za reprezentativniji skup ovih podataka je uputno povećati broj uzoraka u kalibracijskom skupu, ali to je izvan teme ovoga doktorskog rada.

Nadalje, bilo je potrebno idetificirati prisutnost netipičnih i ekstremnih uzoraka u kalibracijskom skupu PMPS A uzoraka, kako bi se formirao robustan PCA model, dobrih izvedbenih sposobnosti na skupu reprezentativnih uzoraka (Slike 157.-166.). Identifikacija netipičnih i ekstremnih uzoraka provedena je pomoću prikaza rezidualne X-varijance i uzoraka visoke utjecajne vrijednosti, koji su ukazali na uzorke koji su potencijalni netipični uzorci i bilo ih je potrebno pažljivo dodatno analizirati. Pri analizi takvih uzoraka istraženo je da li se radi samo o ekstremnim uzorcima, koje želimo zadržati radi realnog prikaza varijabilnosti među uzorcima, ili je ekstremni uzorak ipak netipični uzorak, kojeg je potrebno izuzeti iz kalibracijskog skupa PMPS A.

Klasifikacijski model formiran na skupini uzoraka koji sadrže netipičan uzorak ima loše izvedbene karakteristike i za sam model i za nove objekte, radi pogrešno formirane granice oko skupine u PCA prostoru.

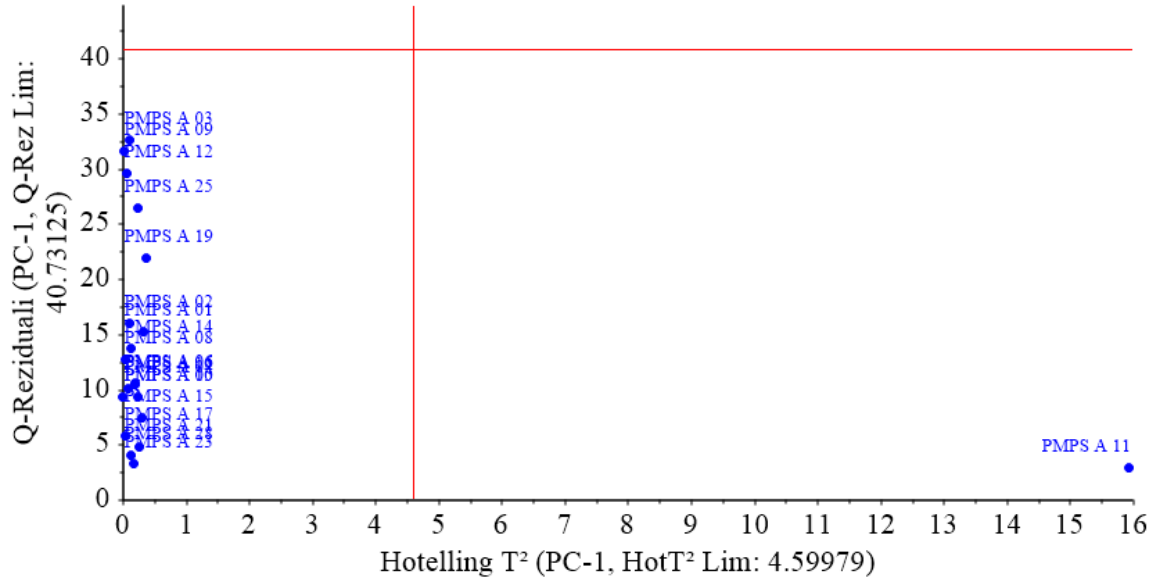


Slika 157. Rezidualne X-varijanca i utjecajna vrijednosta uzoraka PMPS A za PC1.

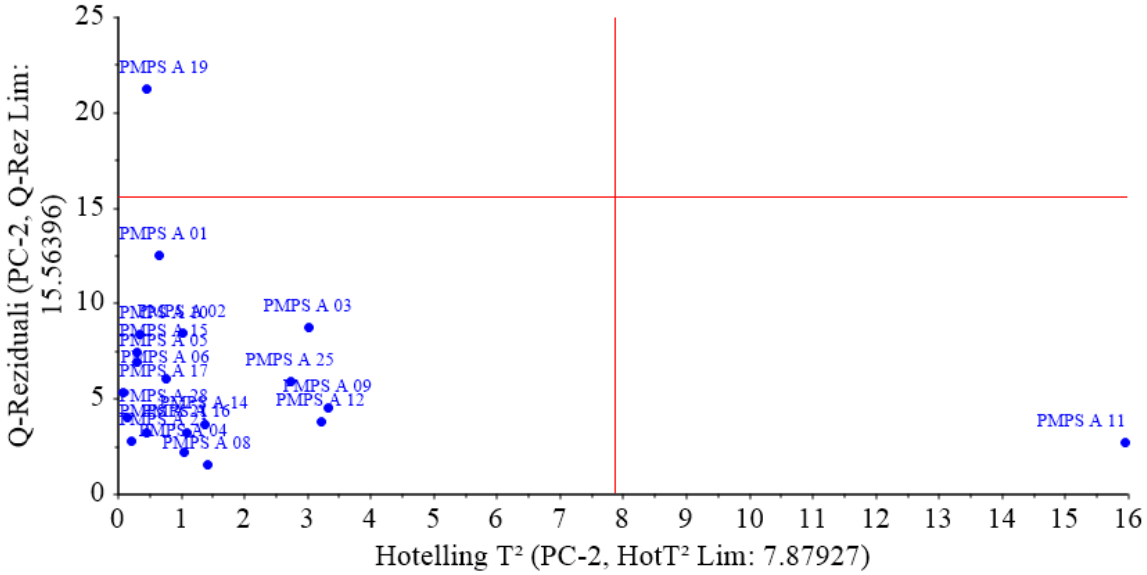


Slika 158. Rezidualna X-varijanca i utjecajna vrijednost uzoraka PMPS A za PC2.

Na Slikama 157. i 158. prikazane su rezidualne X-varijance za svaki uzorak PMPS A prema ostalim uzorcima te utjecajne vrijednosti uzoraka na glavne komponente. Na ovim slikama se može vidjeti da uzorak PMPS A 11 odstupa od ostalih uzoraka (visoki utjecaj na PCA model) i znatno utječe na kvalitetu PCA modela, a to ukazuje da je uzorak PMPS A 11 potencijalni netipični uzorak. Kako bi dodatno istražili utjecaj uzorka PMPS A 11 na formiranje PCA modela, načinjene su dodatne statističke analize i to prikaz Hotelling T^2 statistike i Q-reziduala (Slike 159. i 160.) za prve dvije glavne komponente, zatim prikaz utjecaja PMPS A uzoraka na kalibracijski skup ovih uzoraka PMPS A, odnosno identificirani su uzorci visokih utjecajnih vrijednosti (Slike 161. i 162.) na prve dvije glavne komponente. Također je načinjena Hotelling T^2 statistika (Slike 163. i 164.) i prikazani su Q-reziduali (Slike 165. i 166.).

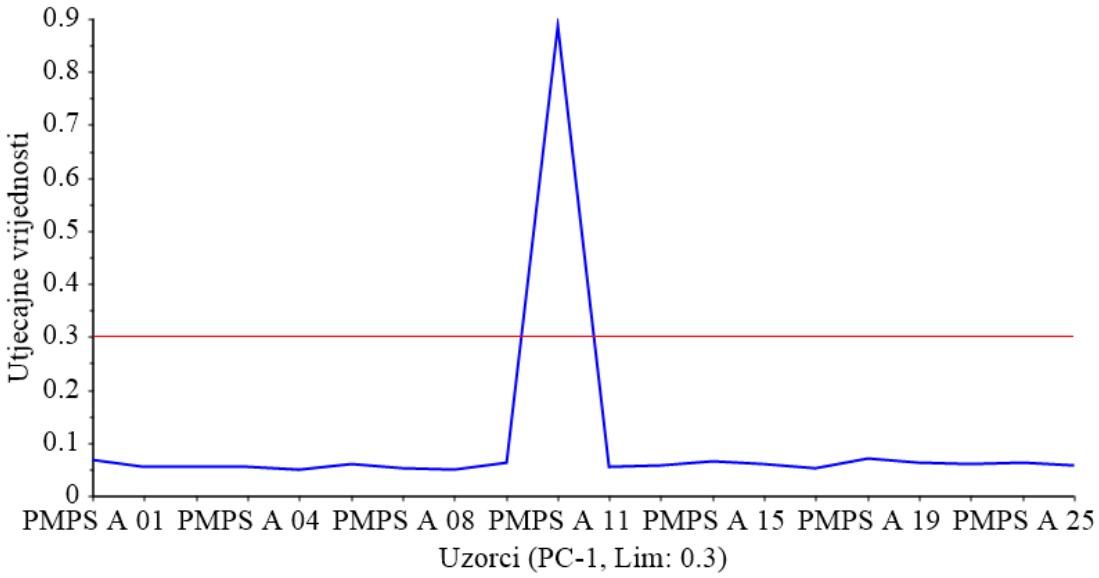


Slika 159. Hotelling T^2 satatistika i Q-reziduali uzoraka PMPS A za PC 1 s pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).

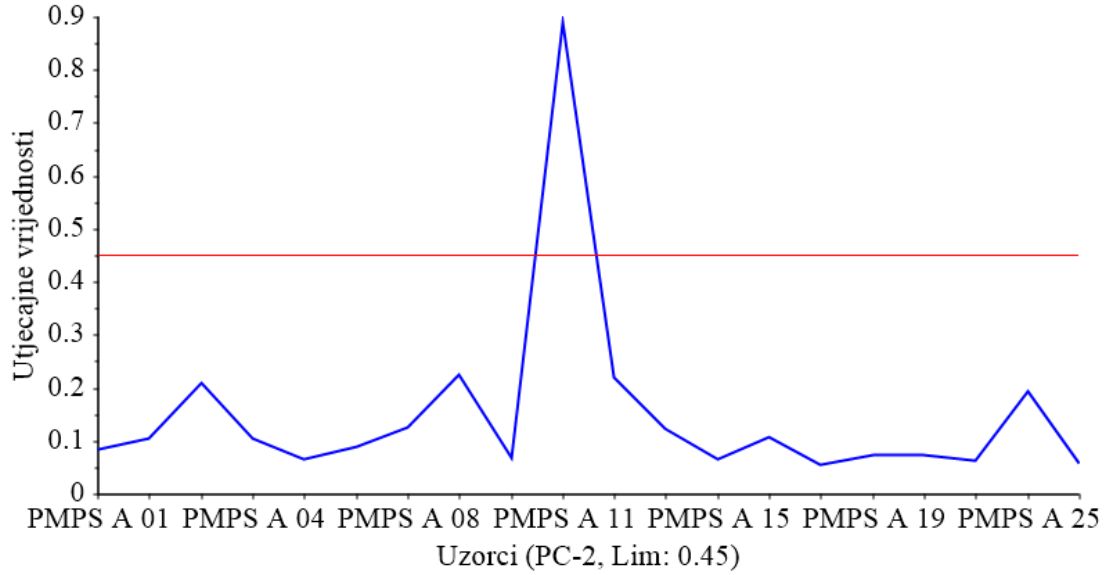


Slika 160. Hotelling T^2 statistika i Q-reziduali uzoraka PMPS A za PC2 s pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).

Na Slikama 159. i 160. uzorak PMPS A 11 je identificiran je kao potencijalni netipični uzorak.

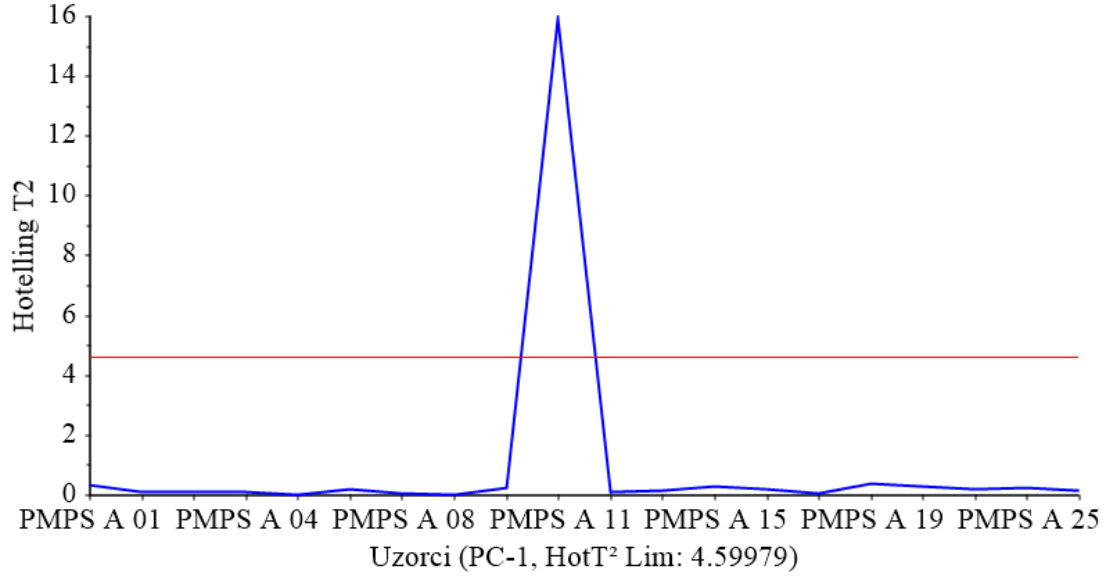


Slika 161. Utjecajne vrijednosti uzoraka PMPS A za PC1 sa pripadajućom kritičnom vrijednosti (crvena linija).

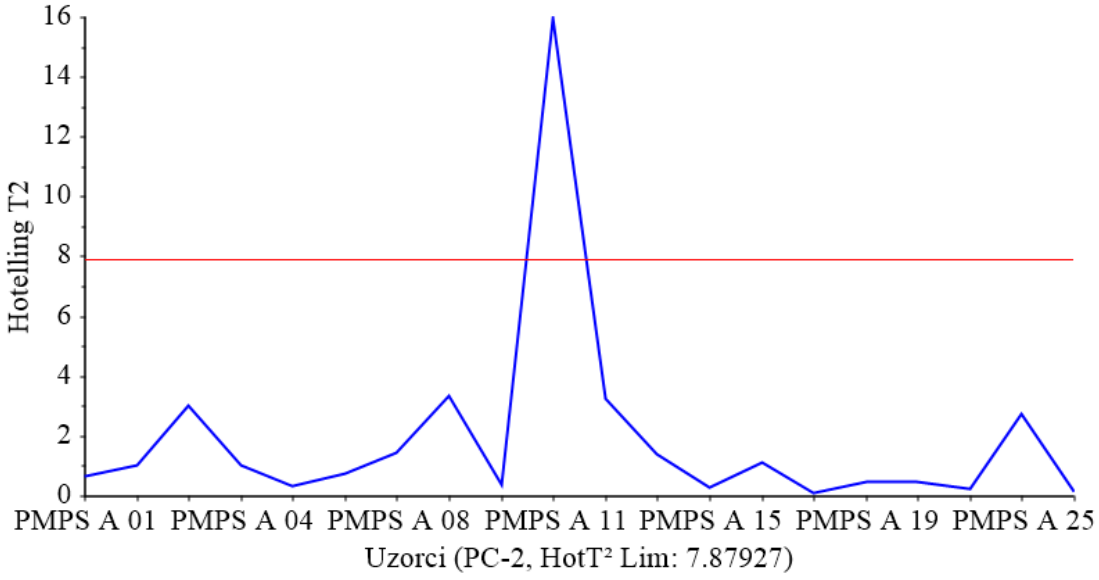


Slika 162. Utjecajne vrijednosti uzoraka PMPS A za PC2 sa pripadajućom kritičnom vrijednosti (crvena linija).

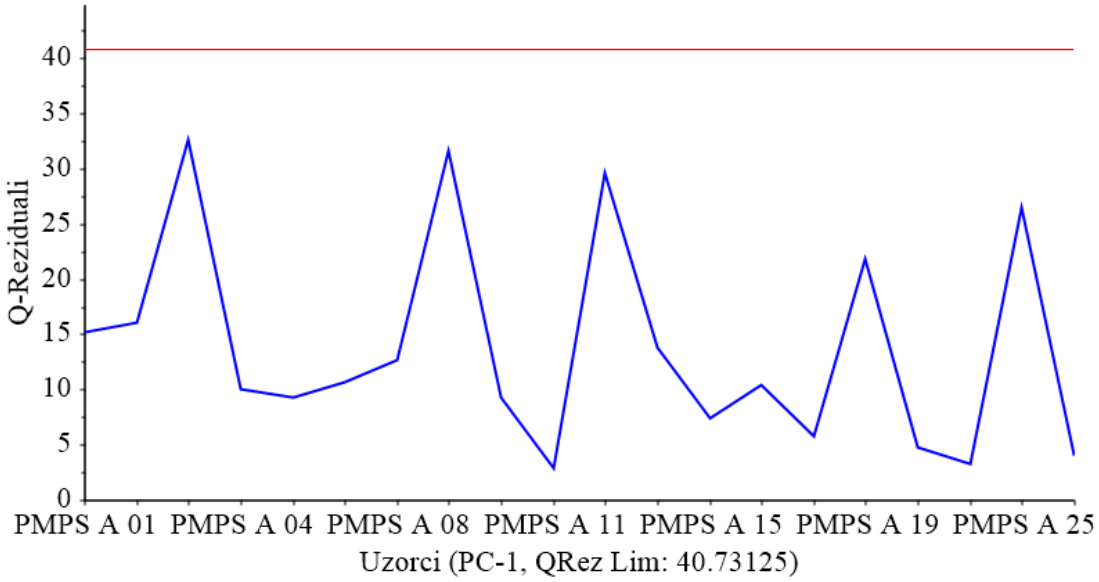
Na Slikama 161. i 162. jasno se može vidjeti kako se uzorak PMPS A 11 znatno razlikuje od preostalih uzoraka PMPS A tj. ima izrazito visoki utjecaj na formiranje PCA modela, odnosno budućeg Raman SIMCA modela.



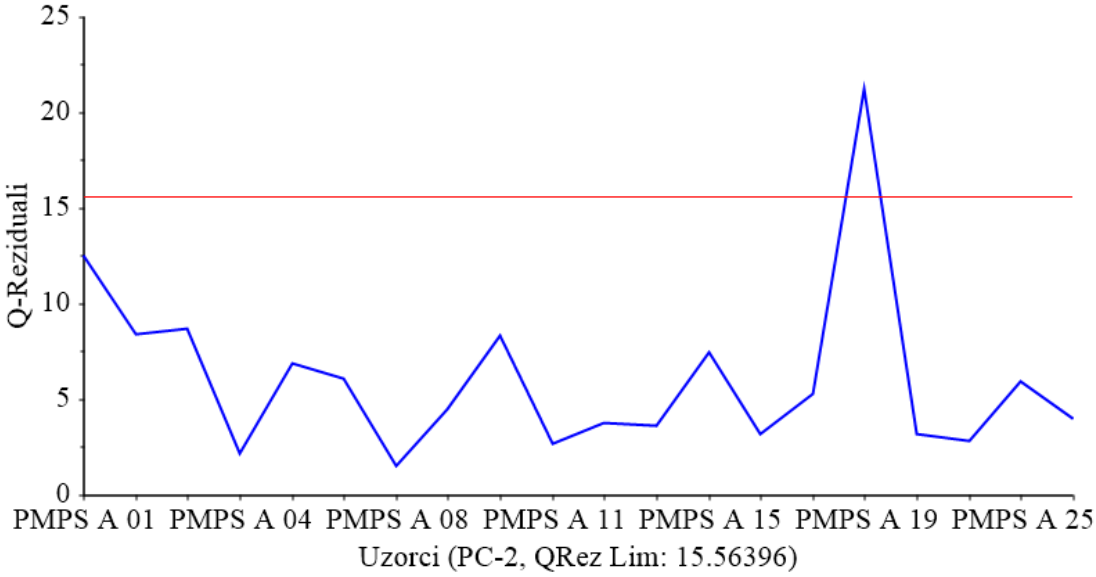
Slika 163. Hotelling T^2 statistika uzoraka PMPS A za PC1 sa pripadajućom kritičnom vrijednosti (crvena linija).



Slika 164. Hotelling T^2 statistika uzoraka PMPS A za PC2 sa pripadajućom kritičnom vrijednosti (crvena linija).



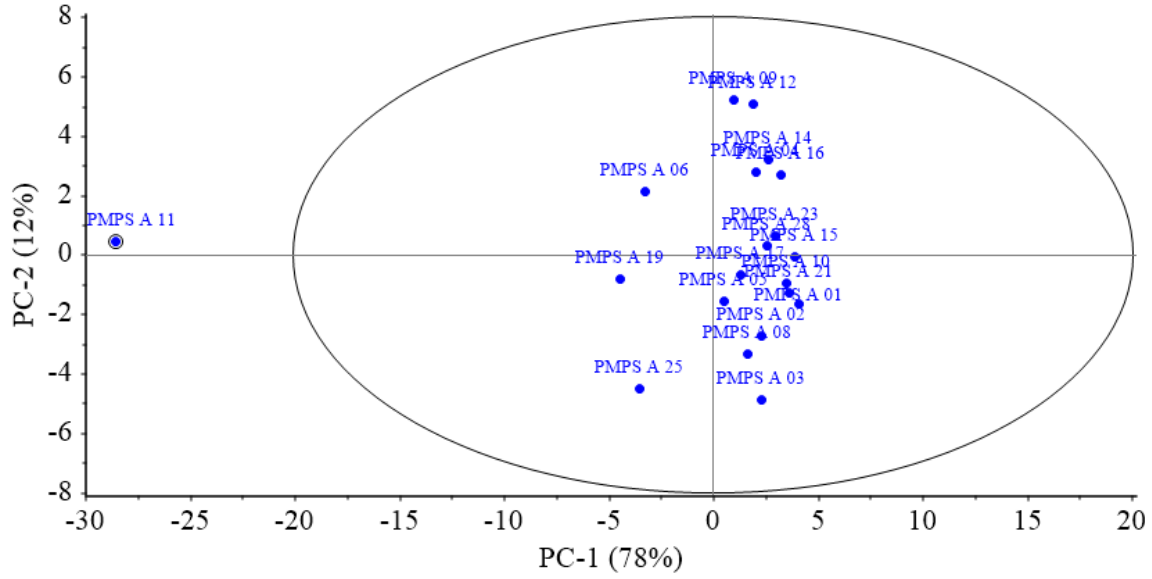
Slika 165. Q-reziduali uzoraka PMPS A za PC1 s pripadajućom graničnom linijom (crvena linija).



Slika 166. Q-reziduali uzoraka PMPS A za PC2 s pripadajućom graničnom linijom (crvena linija).

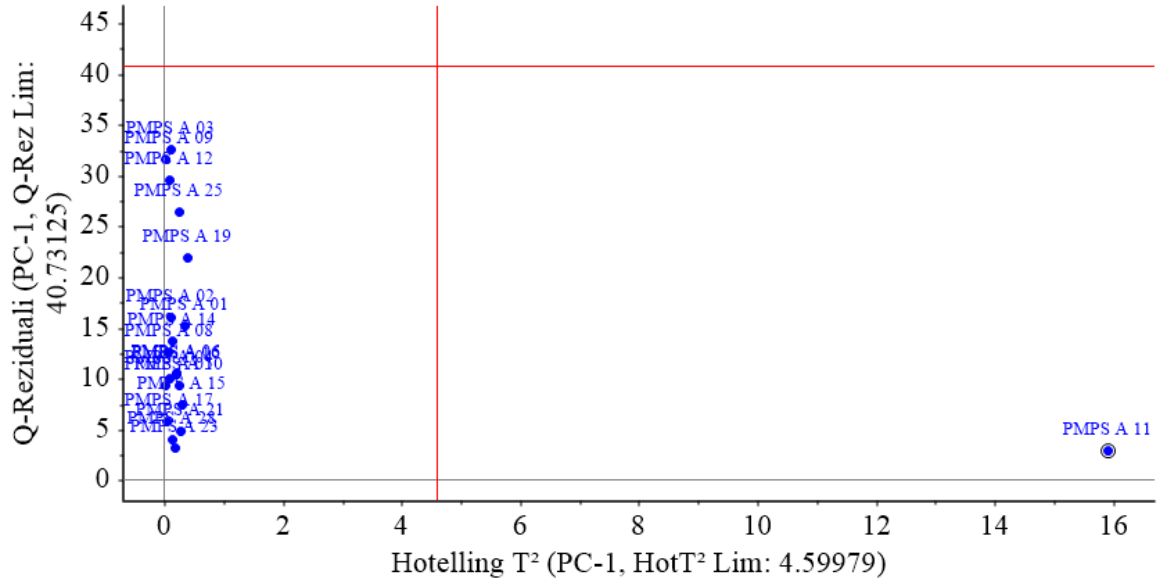
Sukladno rezultatima prikazanim na Slikama 163. i 164. za Hotelling T^2 statistiku, nedvojbeno je da je uzorak PMPS A 11 netipični uzorak i potrebno ga je izuzeti iz kalibracijskog skupa uzoraka PMPS A. Osim ovoga uzorka, identificiran je i uzorak PMPS A 19 sa pripadajućom vrijednosti Q-reziduala znatno većom od preostalih PMPS A uzoraka (Slika 166.). Q-reziduali ukazuju na udaljenost uzorka do modela te pokazuju koliko dobro je uzorak u skladu s modelom. Q statistika je osjetljivija od T^2 statistike i svaka manja promjena u karakteristikama sustava je uočljiva, dok T2 ima veću varijancu i zahtjeva veliku promjenu u karakteristikama sistema kako bi se uočila. Iz gornjih rezultata sasvim je jasno da uzorak PMPS A 19 nije dobro opisan modelom. Naknadnim pregledom svih rezultata i na temelju statističke analize je odlučeno da se uzorak PMPS A 19 smatra ekstremnim uzorkom te se nije izdvojio iz kalibracijskog skupa PMPS A uzoraka.

Nadalje, na temelju eksperimentalnih podataka i statističke analize Raman spektralnih podataka za PMPS A, utvrđeno je da je svakako potrebno izuzeti uzorak PMPS A 11 kao netipičan uzorak. Zbog toga je bilo potrebno ponoviti formiranje PCA modela i to na temelju Raman spektralnih podataka kalibracijskog seta, ali bez ovog netipičnog (PMPS A 11) uzorka. Slika 167. prikazuje netipični uzorak, kojeg treba izuzeti iz skupa podataka za PMPS A.



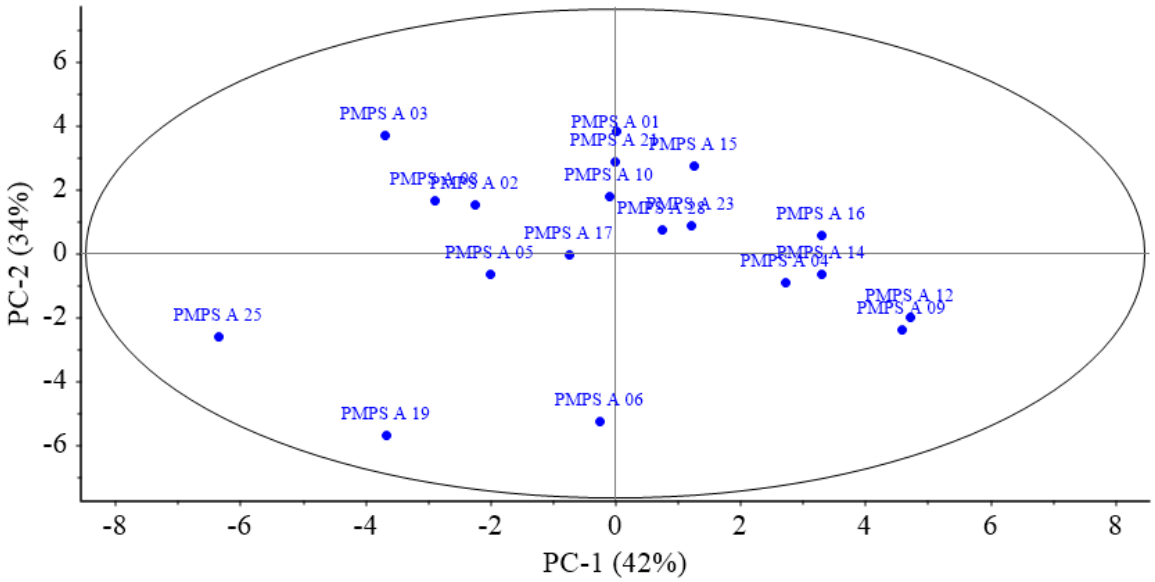
Slika 167. Raspodjela faktorskih bodova PC1 i PC2 matematički obrađenih (SNV i uklanjanje trenda polinomom 4. stupnja) Raman spektara PMPS A u spektralnom području $\tilde{\nu}$ = 3500 - 100 cm^{-1} sa Hotelling T^2 elipsom (interval pouzdanosti 95 %) i označenim netipičnim uzorkom.

Ovaj netipični uzorak istaknut je i na Slici 168. (ovdje ispod).

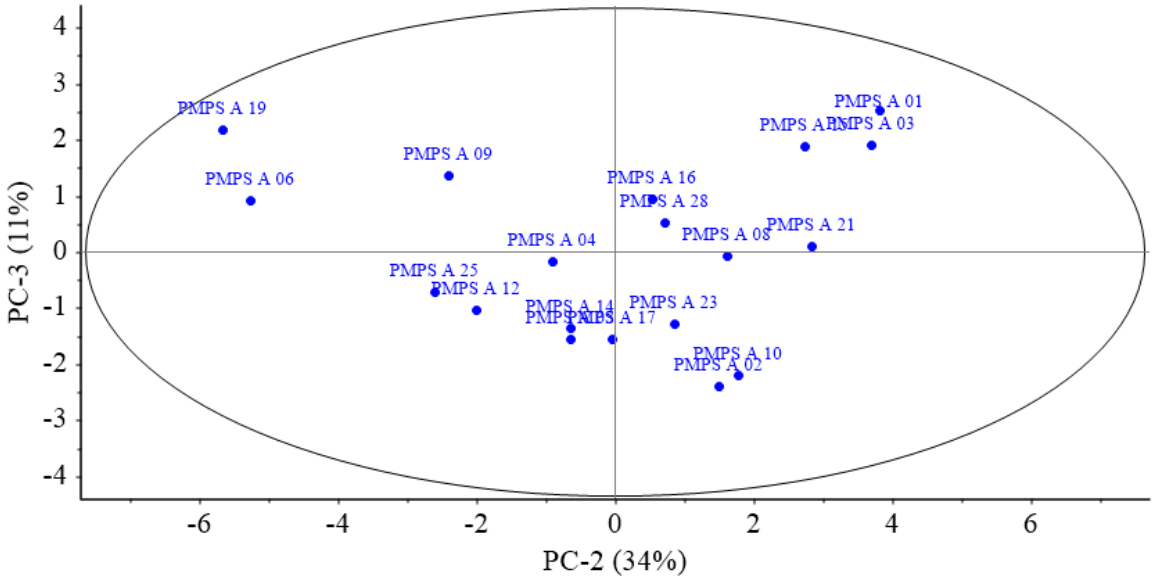


Slika 168. Hotelling T^2 statistika i Q-reziduali uzoraka PMPS A za PC1 s označenim netipičnim uzorkom i pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).

Ponovljena PCA statistička analiza Raman spektralnih podataka PMPS A nakon isključivanja netipičnog uzorka (PMPS A 11) prikazana je na donjoj slici.

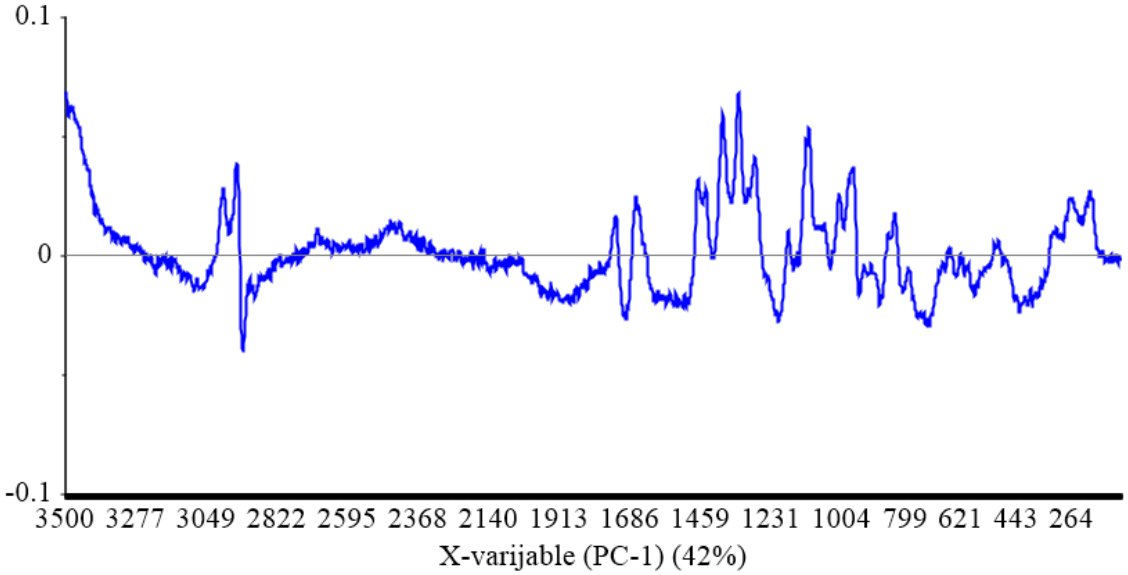


Slika 169. Raspodjela faktorskih bodova PC1 i PC2 matematički obrađenih (SNV i uklanjanje trenda polinomom 4. stupnja) Raman spektara PMPS A u spektralnom području $\tilde{\nu}$ = 3500 - 100 cm^{-1} sa Hotelling T^2 elipsom (interval pouzdanosti 95 %) nakon uklanjanja netipičnog uzorka.

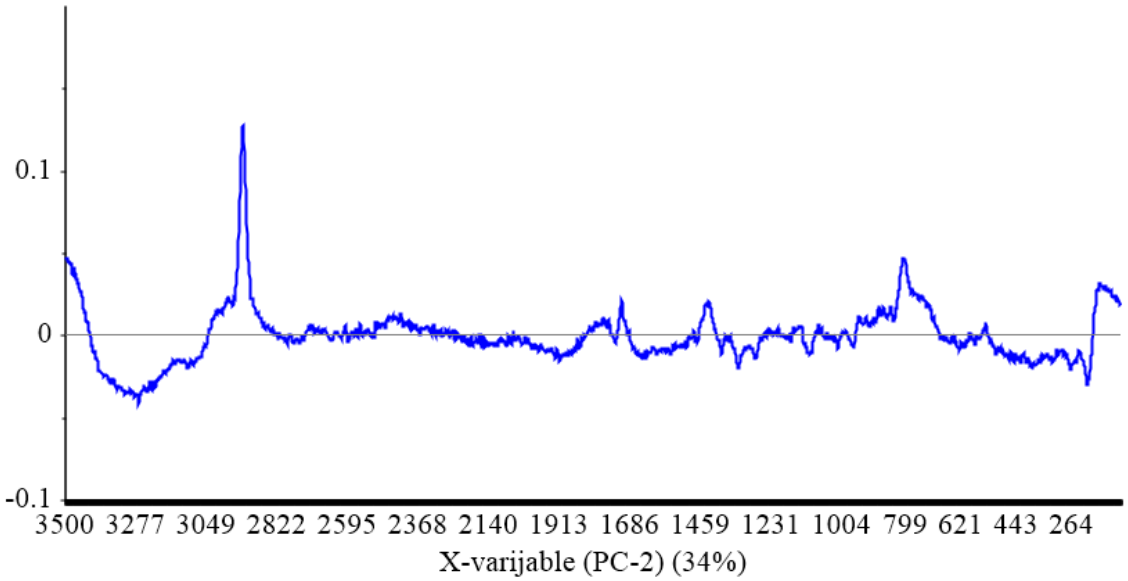


Slika 170. Raspodjela faktorskih bodova PC2 i PC3 matematički obrađenih (SNV i uklanjanje trenda polinomom 4. stupnja) Raman spektara PMPS A u spektralnom području $\tilde{\nu}$ = 3500 - 100 cm^{-1} sa Hotelling T^2 elipsom (interval pouzdanosti 95 %) nakon uklanjanja netipičnog uzorka.

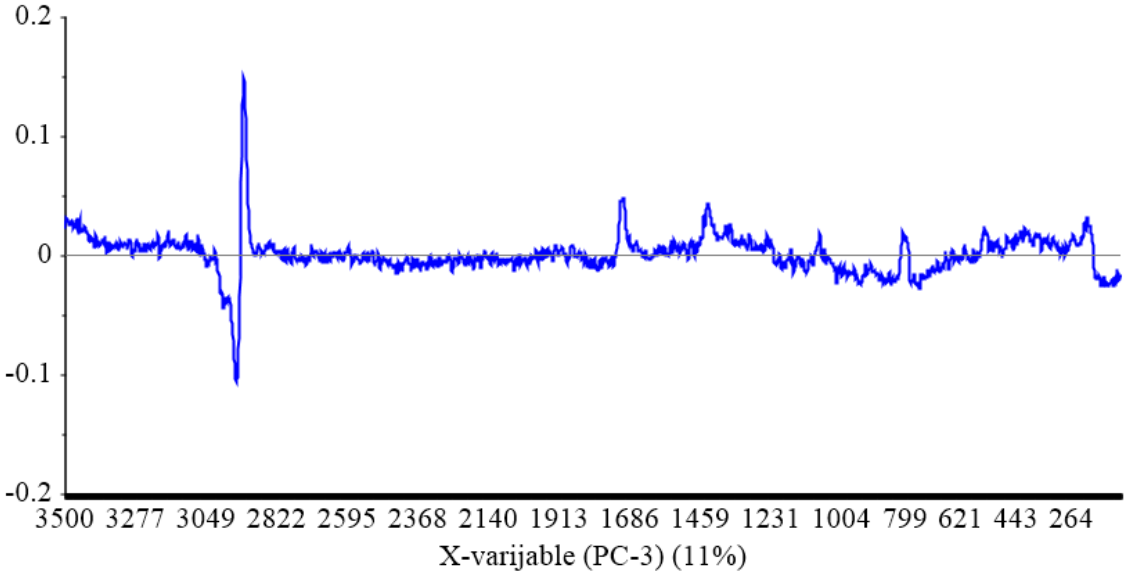
Slike 169. i 170. prikazuju jednoliko raspodjeljene PMPS A uzorke kroz cijelo područje. Linijska opterećenja koja su prikazana na Slikama 171.-174. (ovdje ispod) prikazuju najutjecajnije varijable na pojedinu glavnu komponentu PCA modela, odnosno varijable koje imaju najveći utjecaj na model. Varijable s velikim opterećenjima su varijable koje najviše variraju, te su odgovorne za razliku među uzorcima.



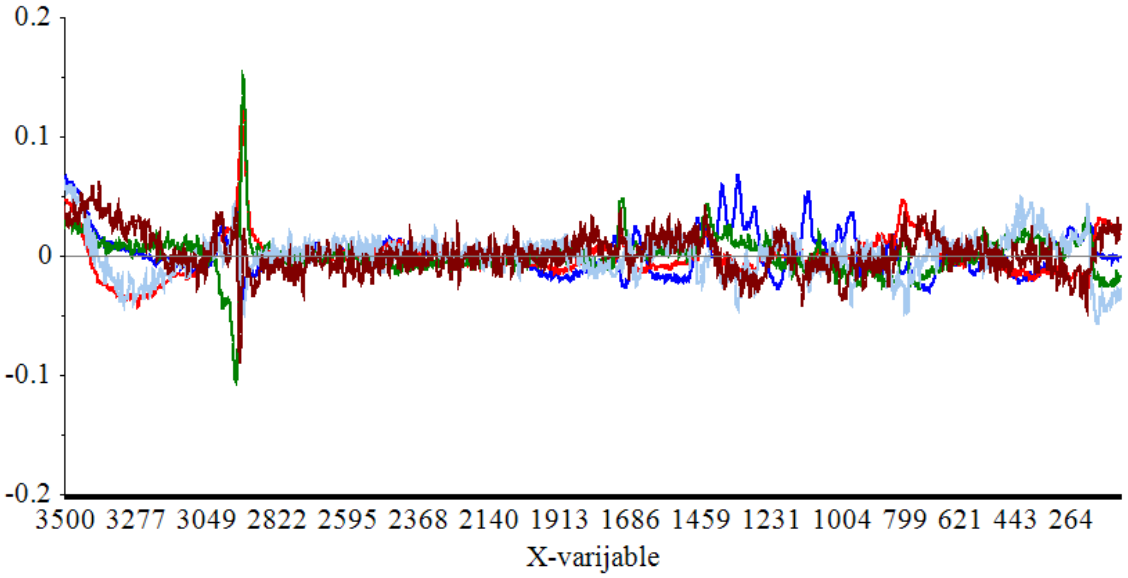
Slika 171. PC1 opterećenja po valnim brojevima ($\tilde{\nu}$) dobivena PCA analizom Raman spektara PMPS A.



Slika 172. PC2 opterećenja po valnim brojevima ($\tilde{\nu}$) dobivena PCA analizom Raman spektara PMPS A.



Slika 173. PC3 opterećenja po valnim brojevima ($\tilde{\nu}$) dobivena PCA analizom Raman spektara PMPS A.

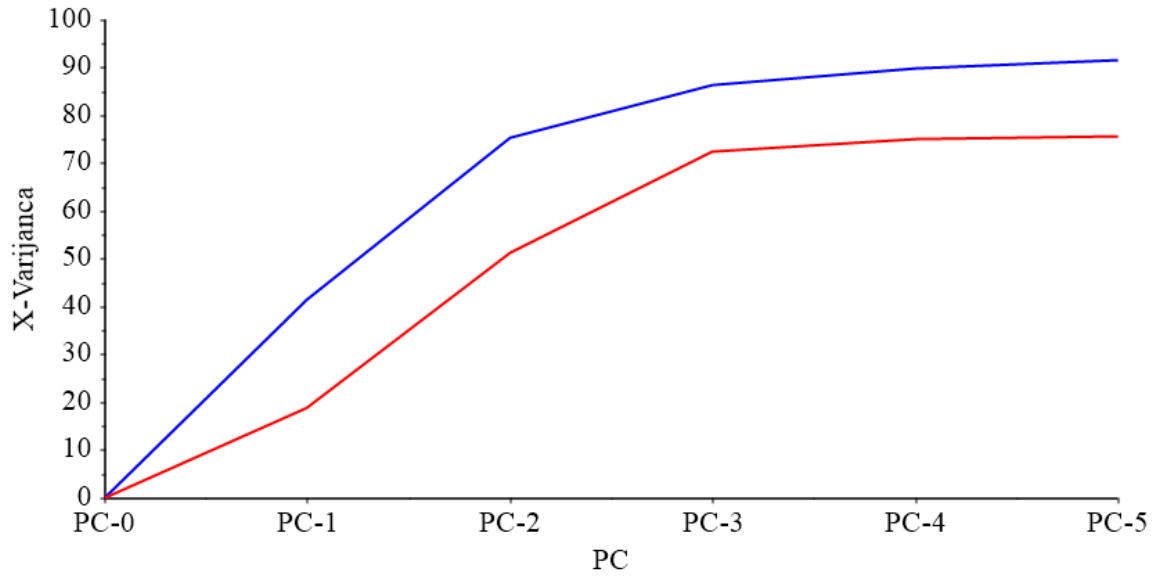


Slika 174. PC1 (plavo), PC2(crveno), PC3 (zeleno), PC4 (sivo) i PC5 (smeđe) opterećenja po valnim brojevima ($\tilde{\nu}$) dobivena PCA analizom Raman spektara PMPS A.

Na slikama 171 - 174. može se vidjeti najutjecajnije varijable na pojedinu glavnu komponentu u formiranju PMPS A modela u regijama oko $\tilde{\nu}$ = 2940 cm^{-1} proizlaze od CH_2 asimetričnog istezanja; $\tilde{\nu}$ = 1750 - 1630 cm^{-1} proizlaze od C=ONHR vibracija istezanja; $\tilde{\nu}$ = 1500 - 1200 cm^{-1} C-H/CH_2 deformacijske vibracije; $\tilde{\nu}$ = 1350 - 1140 proizlaze od P=O istezanja; $\tilde{\nu}$ = 1200 -

950 cm^{-1} proizlaze od C-C i C-O simetričnog istezanja; $\tilde{\nu}$ = 1150 - 1070 cm^{-1} proizlazi od C-O-C asimetričng istezanja; $\tilde{\nu}$ = 980 - 965 cm^{-1} proizlaze od C-H savijanja van ravnine; $\tilde{\nu}$ = 800 - 100 deformacijske vibracije C-C-O grupa (Larkin, 2018).

Nakon ponovljene PCA analize Raman spektralnih podataka za PMPS A, načinjen je i prikaz kumulativne kalibracijske i validacijske varijance (Slika 175. i Tablica 10.), kako bi se odredila složenost projekcijskog modela, tj. utvrdio optimalni broj PC-ova za Raman PCA model PMPS A.



Slika 175. Kumulativna kalibracijska (plava) i validacijska (crvena) varijanca za svaki PC nakon uklanjanja netipičnog uzorka.

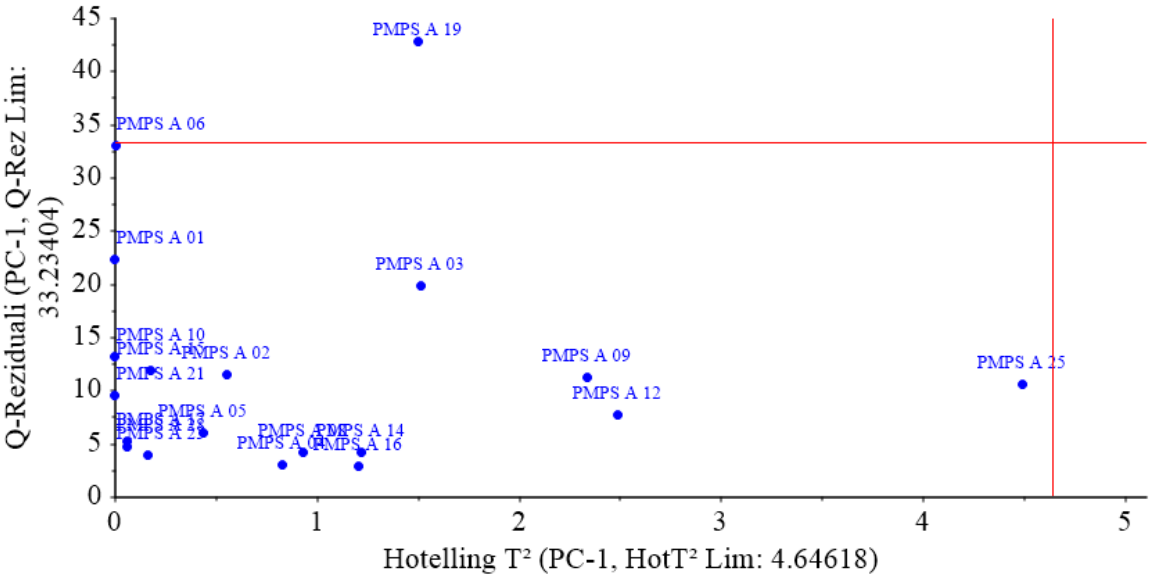
Tablica 10. Kumulativna kalibracijska i validacijska varijanca za svaki PC nakon uklanjanja netipičnog uzorka.

	PC0	PC1	PC2	PC3	PC4	PC5
Kalibracija	0	41.5215	75.3345	86.2948	89.9954	91.6818
Validacija	0	18.9311	51.3027	72.5877	75.1772	75.7426

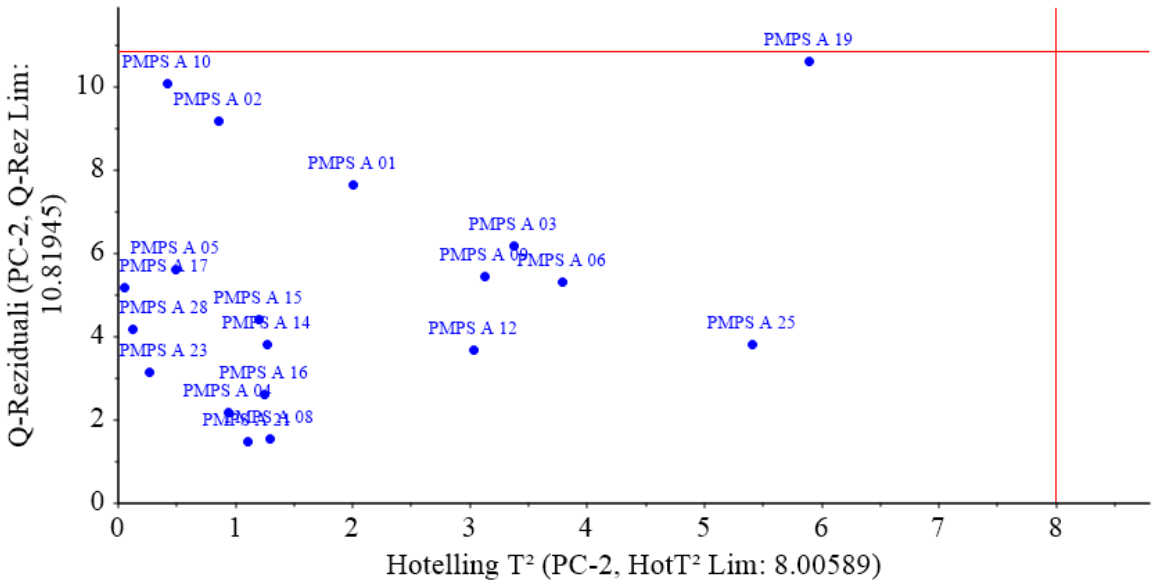
Iz Tablice 10. se može vidjeti da tri PC-a (PC1, PC2 i PC3) obuhvaćaju više od 86 % ukupne kalibracijske varijance, dok četiri PC-a ne obuhvaćaju značajno više (90 %) ukupne varijance. Odabirom tri PC-a umjesto četiri PC-a za model, dobiva se jednostavniji klasifikacijski model

što je poželjno u smislu stabilnosti i interpretacije modela. Također, odabirom većeg broja PC-ova klasifikacijski model nebi imao značajno bolje izvedbene karakteristike, pa je radi svega navedenog, odabrani broj PC-ova za ovaj PCA model tri (3 PC-a). Također se može vidjeti i da tri PC-a (PC1, PC2 i PC3) obuhvaćaju više od 72 % ukupne validacijske varijance, dok preostali PC-ovi obuhvaćaju neznatno više (ukupno 75 %) ukupne varijance. Može se smatrati da razlika između kalibracijske i validacijske varijance potječe od relativno malog broja uzoraka kalibracijskog seta te je u budućnosti potrebno povećati broj uzoraka kako bi se dobio što reprezentativniji set uzoraka za formiranje i optimiranje modela.

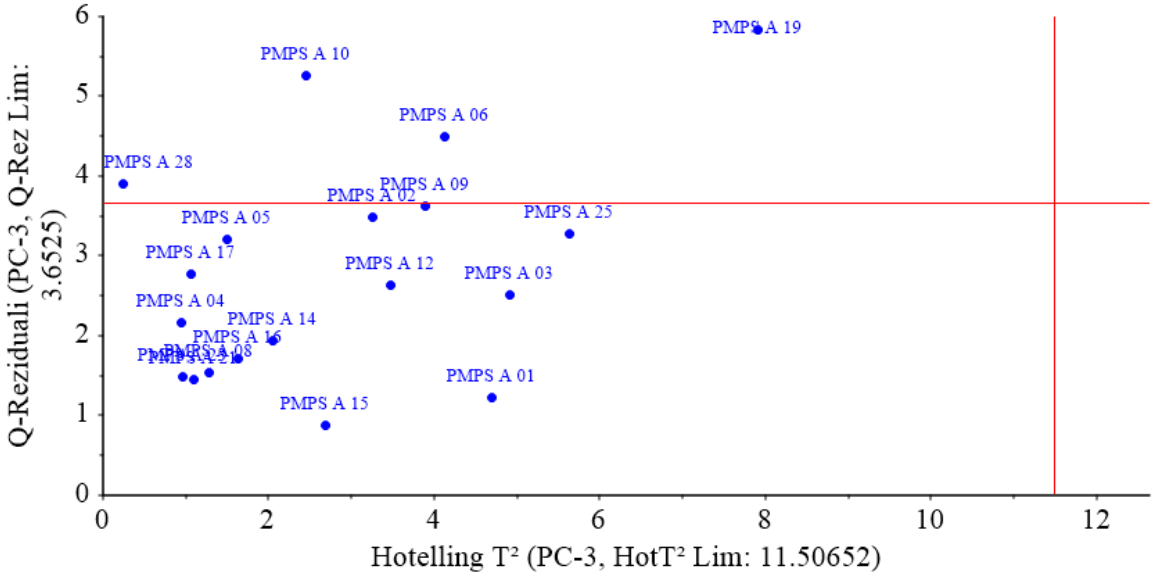
Nadalje, načinjena je potpuna statistička analiza kalibracijskog skupa PMPS A podataka kako bi se identificirali eventualno prisutni dodatni netipični ali i ekstremni uzorci (ovdje ispod).



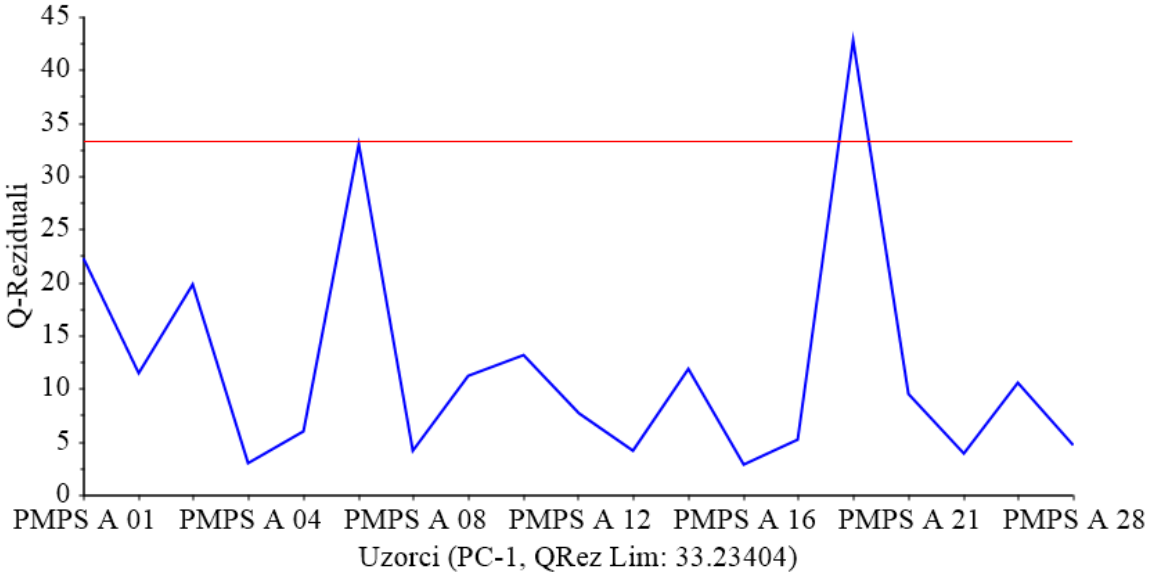
Slika 176. Hotelling T^2 statistika i Q-reziduali uzoraka PMPS A za PC1 s pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).



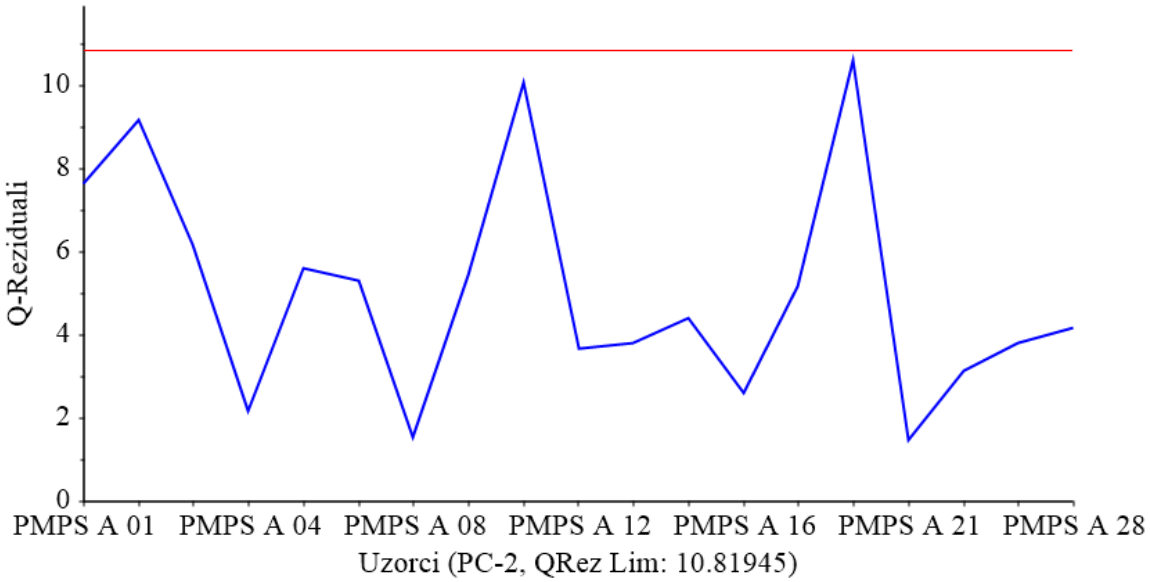
Slika 177. Hotelling T^2 statistika i Q-reziduali uzoraka PMPS A za PC2 s pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).



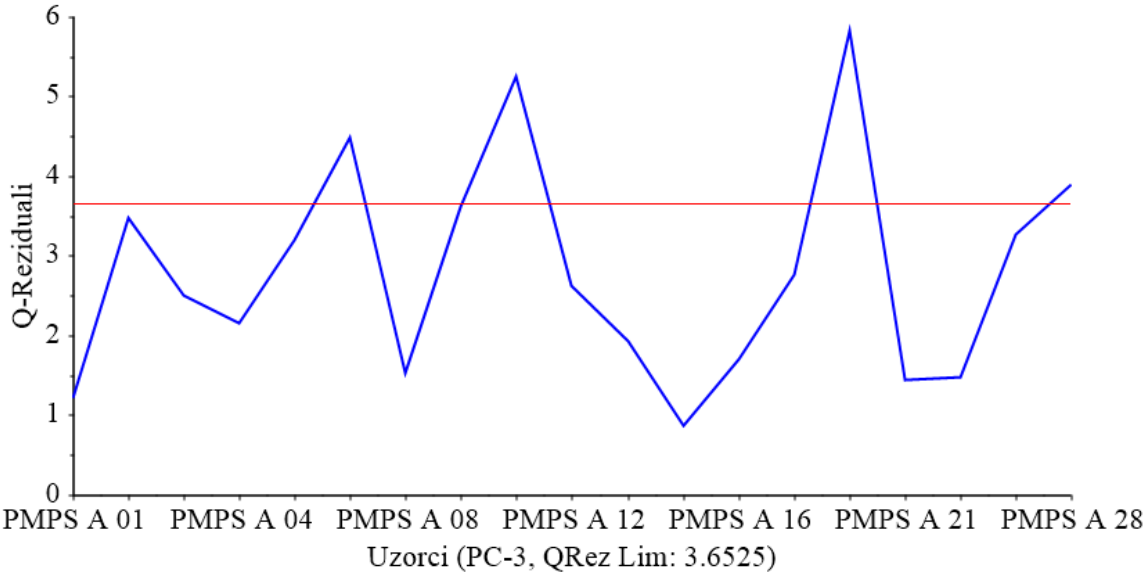
Slika 178. Hotelling T^2 statistika i Q-reziduali uzoraka PMPS A za PC3 s pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).



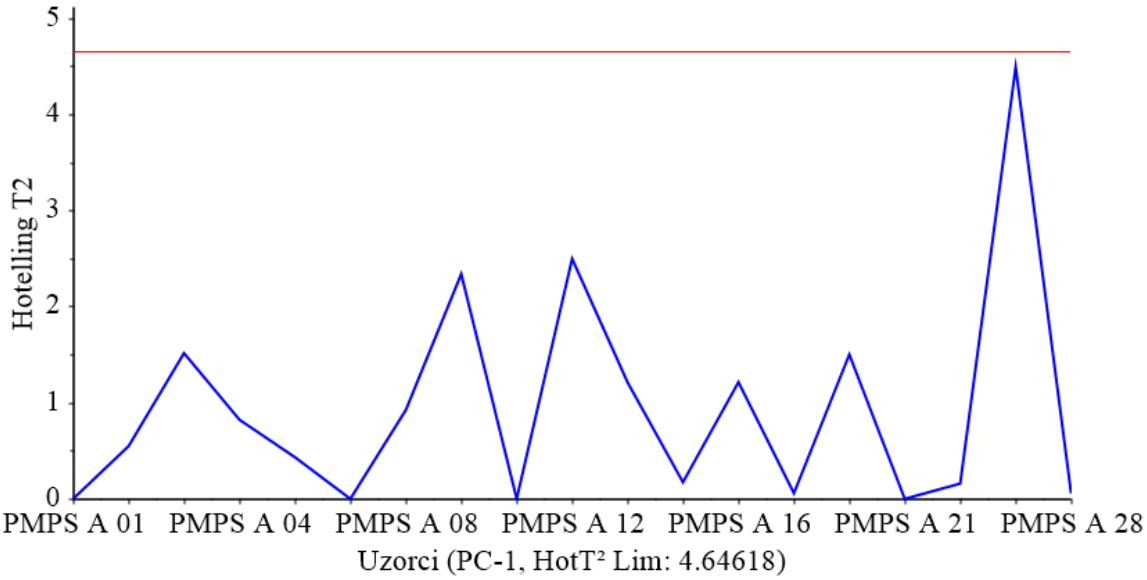
Slika 179. Q-reziduali uzoraka PMPS A za PC1 s pripadajućom graničnom linijom (crvena linija).



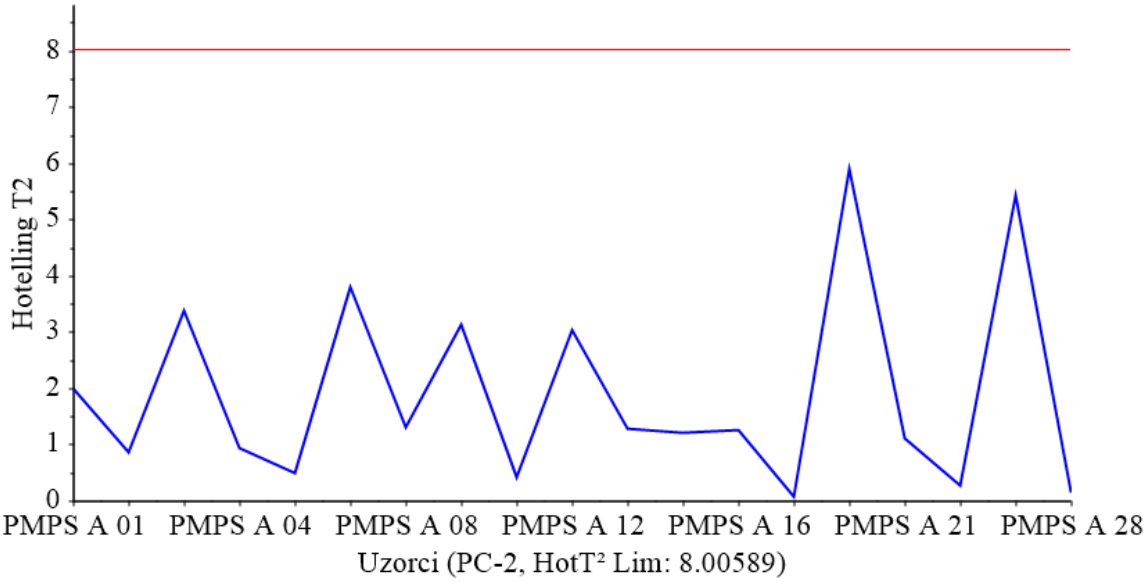
Slika 180. Q-reziduali uzoraka PMPS A za PC2 s pripadajućom graničnom linijom (crvena linija).



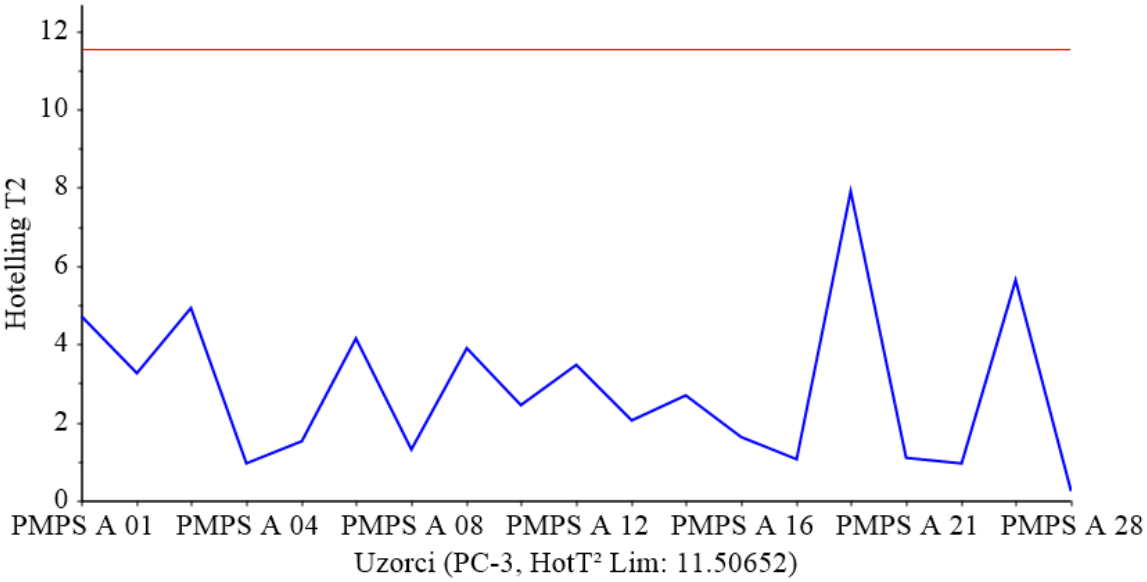
Slika 181. Q-reziduali uzoraka PMPS A za PC3 s pripadajućom graničnom linijom (crvena linija).



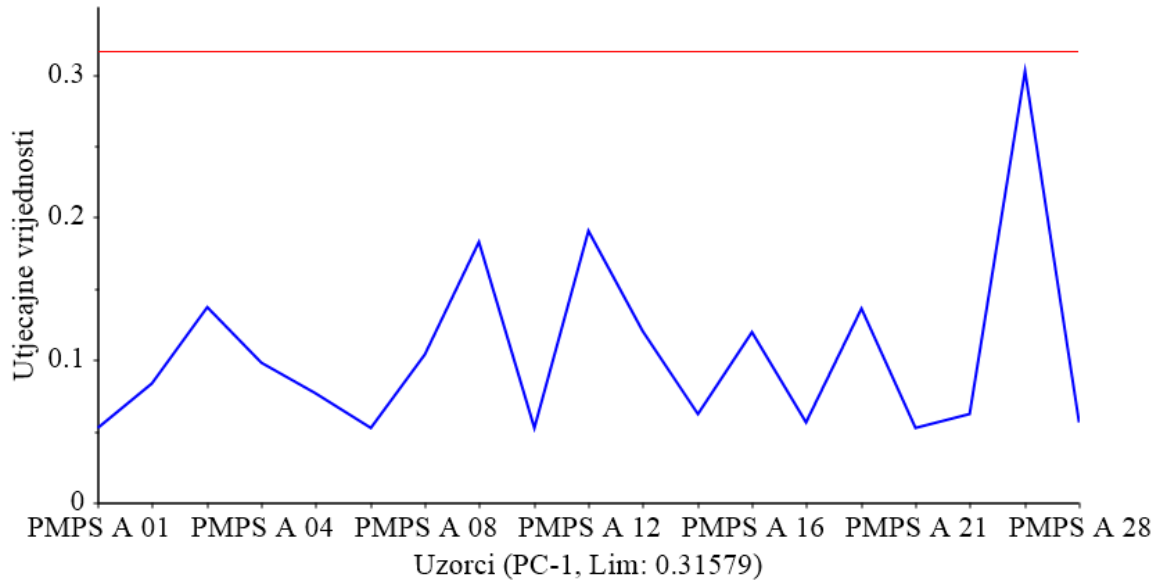
Slika 182. Hotelling T^2 statistika uzoraka PMPS A za PC1 sa pripadajućom kritičnom vrijednosti (crvena linija).



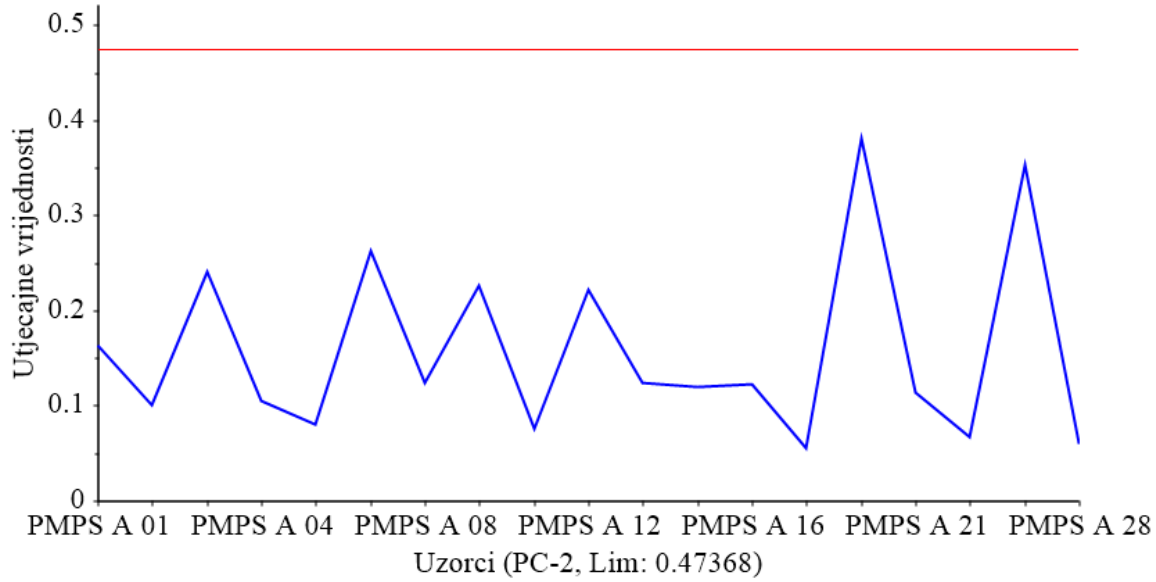
Slika 183. Hotelling T^2 statistika uzoraka PMPS A za PC2 sa pripadajućom kritičnom vrijednosti (crvena linija).



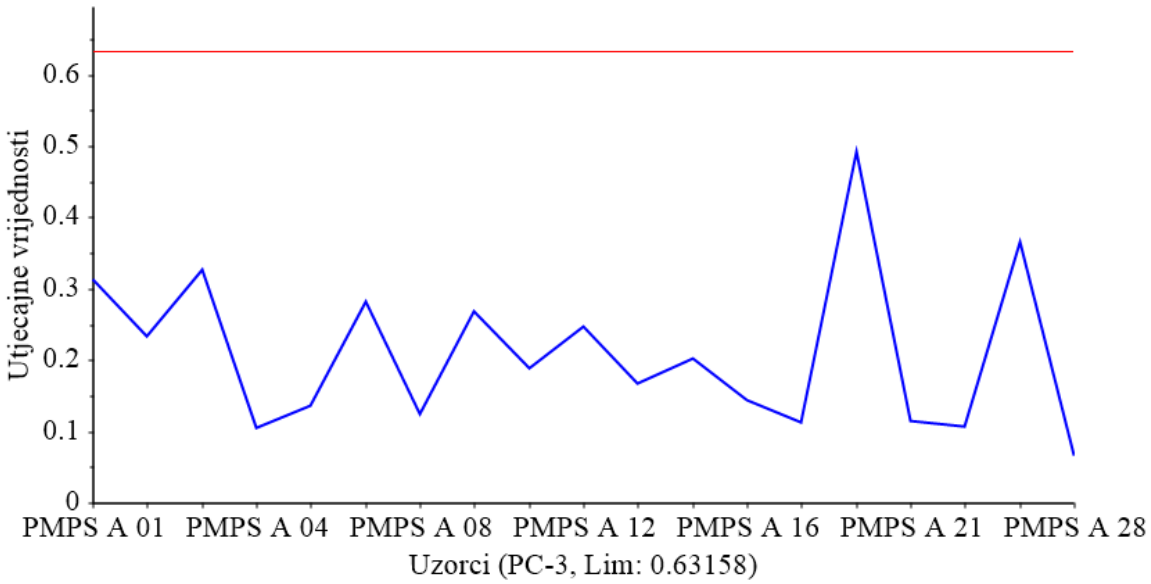
Slika 184. Hotelling T^2 statistika uzoraka PMPS A za PC3 sa pripadajućom kritičnom vrijednosti (crvena linija).



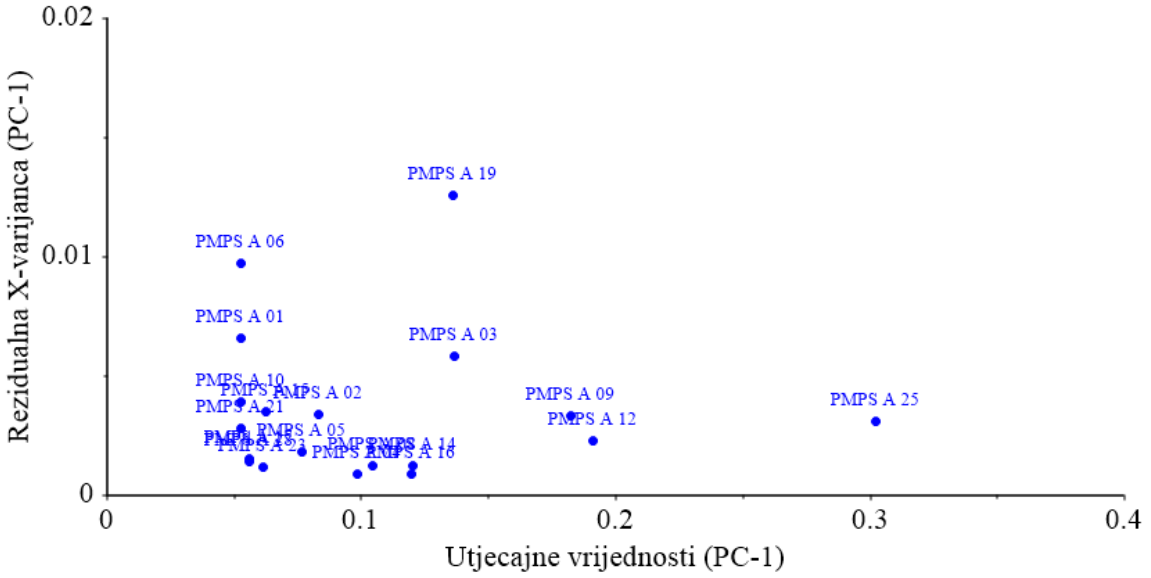
Slika 185. Utjecajne vrijednosti uzoraka PMPS A za PC1 sa pripadajućom kritičnom vrijednosti (crvena linija).



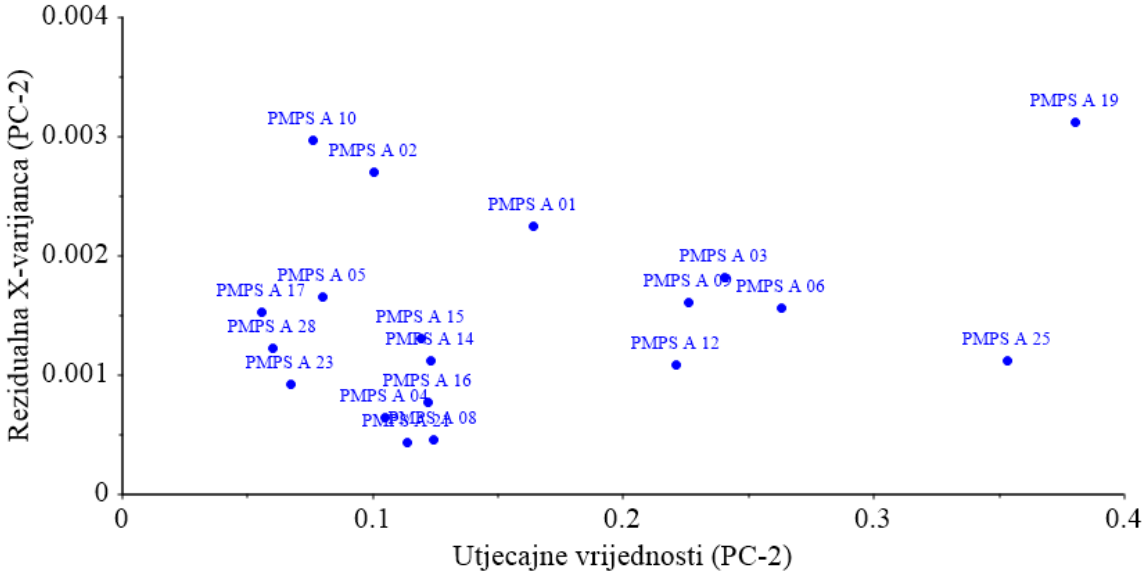
Slika 186. Utjecajne vrijednosti uzoraka PMPS A za PC2 sa pripadajućom kritičnom vrijednosti (crvena linija).



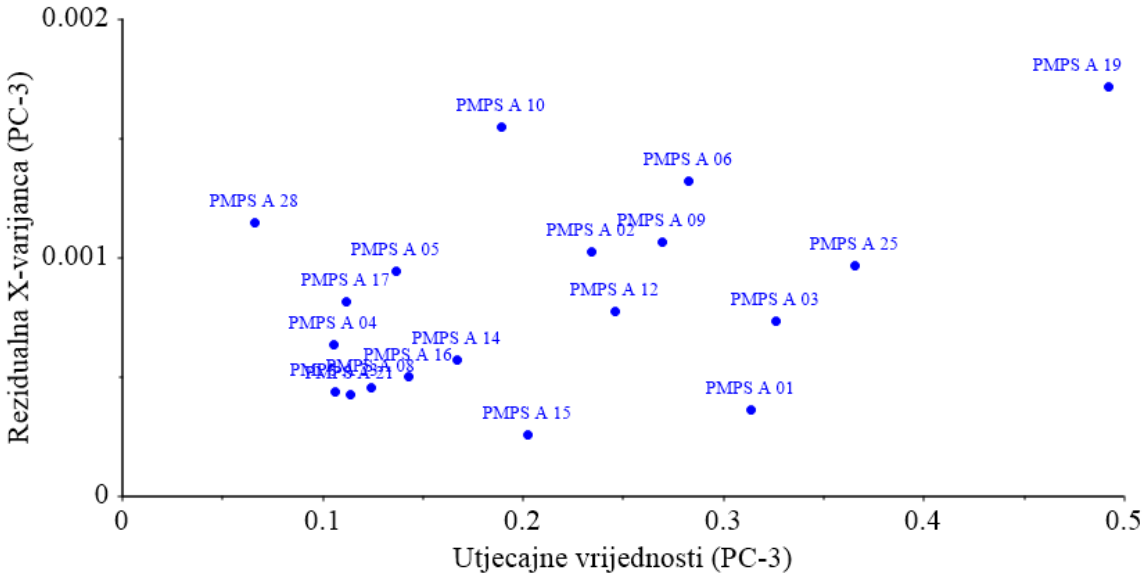
Slika 187. Utjecajne vrijednosti uzoraka PMPS A za PC3 sa pripadajućom kritičnom vrijednosti (crvena linija).



Slika 188. Rezidualna X-varijanca i utjecajna vrijednost uzoraka PMPS A za PC1.



Slika 189. Rezidualna X-varianca i utjecajna vrijednost uzoraka PMPS A za PC2.



Slika 190. Rezidualna X-varijanca i utjecajna vrijednost uzoraka PMPS A za PC3.

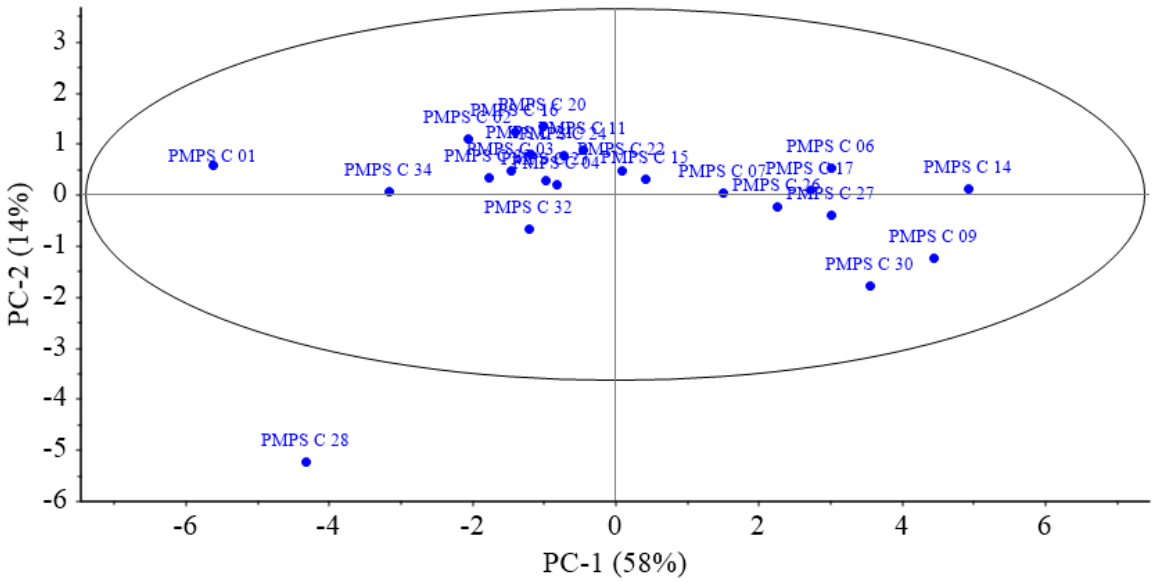
Slike 176. - 190. prikazuju Hotelling T^2 statistiku, Q-reziduale i uzorke visokih utjecajnih vrijednosti. Ovim načinom prikazivanja postiže se bolji uvid utjecaja svakog PMPS A uzorka na PCA model, identificiraju se uzorci udaljeni od središta modela, zatim uzorci koji se razlikuju od prosječnih vrijednosti kao i uzorci sa visokim utjecajem na ovaj model. Na ovaj se način mogu identificirati ekstremni uzorci, koje je onda dodatno potrebno analizirati te po potrebi odbaciti iz ovoga seta.

Na Slikama 176. - 181. se može vidjeti da uzorak PMPS A 19 ima vrijednosti Q-reziduala više od ostalih uzoraka iz ovoga kalibracijskog seta PMPS A uzoraka. Također uzorak PMPS A 19 pokazuje visok utjecaj na model sa visokom rezidualnom X-varijancom (Slike 188 .- 190.). Ovaj je uzorak identificiran i u prethodnom PCA modelu uzoraka PMPS A kao uzorak koji je lošije od ostalih uzoraka opisan PCA modelom te je, nakon provedene dodatne statističke analize, nanovo identificiran kao ekstremni uzorak. Ovaj je uzorak dio dugotrajne stabilitetne studije te različitih fizikalnih karakteristika i kao takav poželjan i jako važan član kalibracijskog skupa uzoraka PMPS A. Iz gornje slike (Slika 181.) se može uočiti da uzorci PMPS A 06, PMPS A 10 i PMPS A 19 imaju više Q-reziduale za PC3. Niti jedan uzorak kalibracijskog seta ne pokazuje visoke vrijednosti Hotelling T^2 statistike (Slike 182. - 184.), niti visoki utjecaj na PCA model (Slike 185. – 187.).

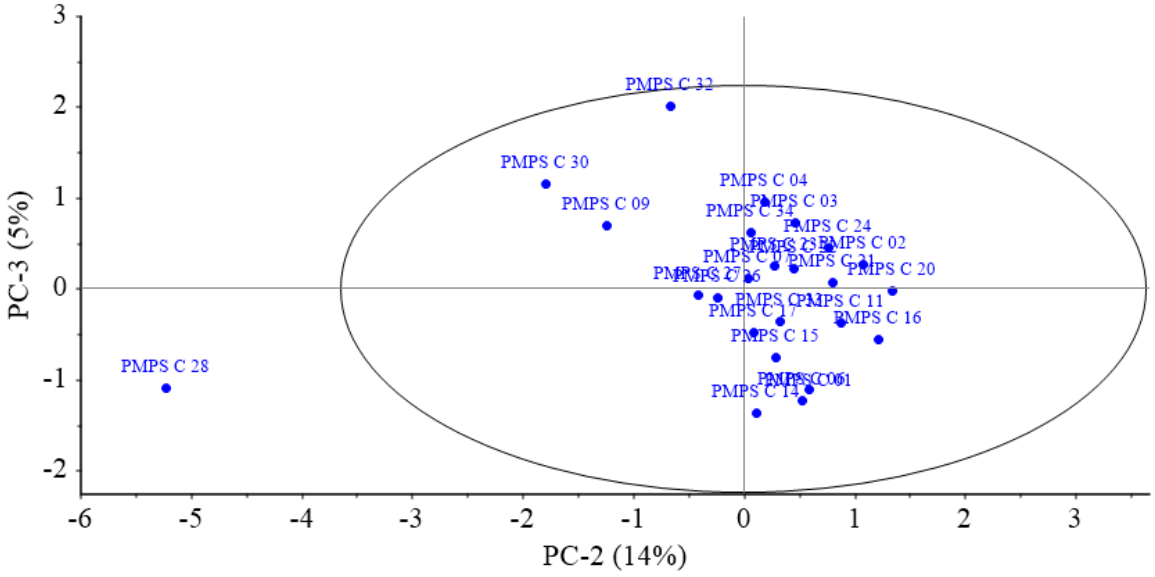
Iz cjelokupne statističke analize može se zaključiti kako su uzorci PMPS A 06, PMPS A 10 i PMPS A 19 ekstremni uzorci koji će se zadržati u kalibracijskom skupu uzoraka, te nije potrebno izdvojiti niti jedan uzorak iz kalibracijskog seta PMPS A.

4.3.5.2. PCA modeliranje Raman spektara PMPS C

Provedena je analiza glavnih komponenata na Raman spektralnim podacima iz kalibracijskog skupa uzoraka PMPS C, koji uključuju 34 Raman spektra. Ovaj kalibracijski skup je također korišten i za optimizaciju PCA modela i to postupkom unakrsne validacije. Slično kao što je provedena analiza glavnih komponenti za PMPS A, ova analiza glavnih komponenti provedena je tako da je načinjena raspodjela faktorskih bodova, zatim opterećenja, te prikaz utjecajnih vrijednosti, kao Hotelling T^2 statistika i Q-reziduali, u cilju identificiranja netipičnih i ekstremnih PMPS C uzoraka.

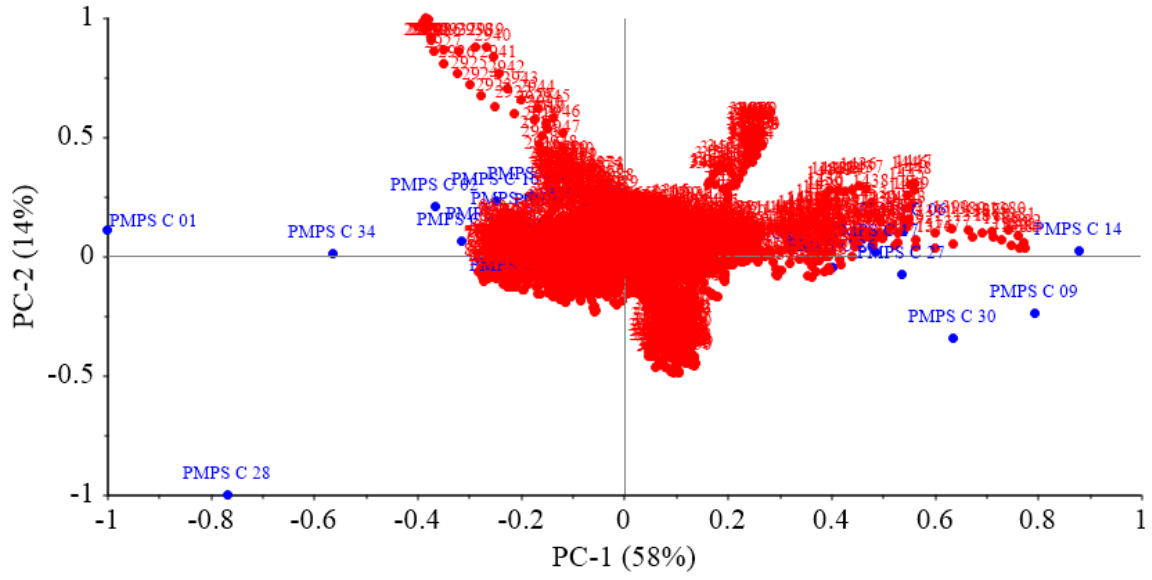


Slika 191. Raspodjela faktorskih bodova PC1 i PC2 matematički obrađenih (SNV i uklanjanje trenda polinomom 4. stupnja) Raman spektara PMPS C u spektralnom području $\tilde{\nu}$ = 3500 - 100 cm^{-1} sa Hotelling T^2 elipsom (interval pouzanosti 95 %) sa centrom modela u sjecištu dvaju pravaca (sivo).



Slika 192. Raspodjela faktorskih bodova PC2 i PC3 matematički obrađenih (SNV i uklanjanje trenda polinomom 4. stupnja) Raman spektara PMPS C u spektralnom području $\tilde{\nu}$ = 3500 - 100 cm^{-1} sa Hotelling T^2 elipsom (interval pouzanosti 95 %) sa centrom modela u sjecištu dvaju pravaca (sivo).

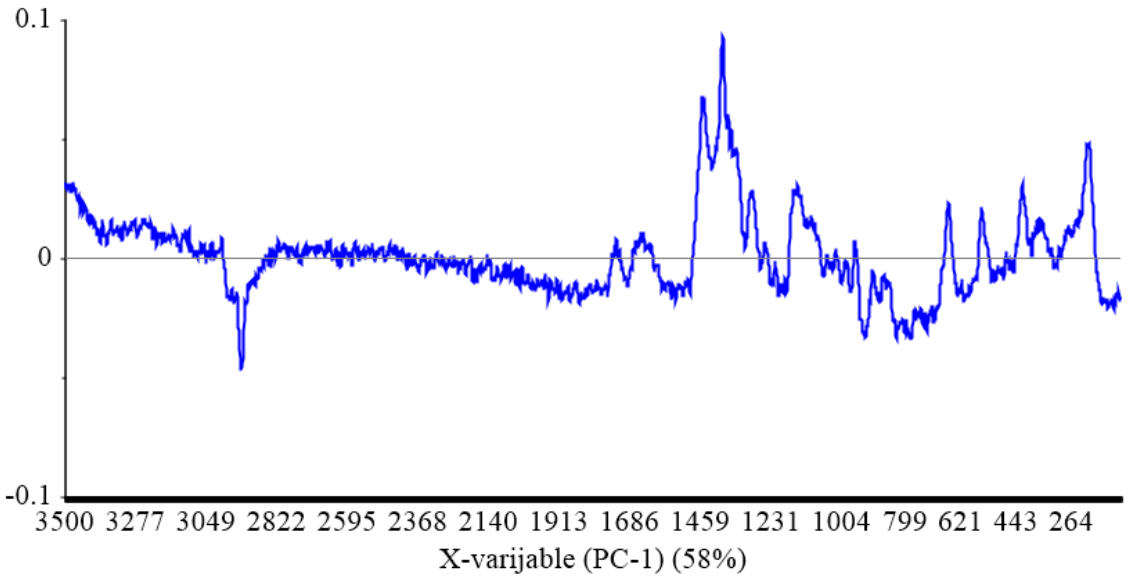
Na Slici 191. su prikazani faktorski bodovi prve i druge glavne komponente, a na Slici 192. prikazani su faktorski bodovi druge i treće glavne komponente. Prva glavna komponenta (PC1) oduhvaća 58 % varijance, druga glavna komponenta (PC2) obuhvaća 14% varijance, dok treća glavna komponenta (PC3) obuhvaća 5 % varijance. Slike 191. i 192. jasno prikazuju jednoliku raspodjelu uzoraka kroz cijelo područije. Nisu prisutni trendovi među uzorcima niti nejednoliko grupiranje ovih uzoraka. Hotelling T^2 elipsa (interval pouzdanosti 95 %) olakšava detekciju netipičnih uzoraka ili ekstremnih PMPS C uzoraka, koje je svakako potrebno dodatno analizirati. Na Slikama 191. i 192. uzorak PMPS C 28 je izvan Hotelling T^2 elipse te ga je potrebno dodatno analizirati statističkim metodama i tako okarakterizirati kao mogući netipični uzorak. Hotelling T^2 varijable su suma n neovisnih Studentovih t varijabli. Kod Hotelling T^2 statistike odgovarajući segment intervala pouzdanosti, ovisno o broju varijabli je omeđen elipsom, elipsoidom ili hiperelipsoidom. Širina segmenta je funkcija varijance, odnosno disperzije varijabli, a kut u odnosu na osi ovisi o stupnji njihove korelacije.



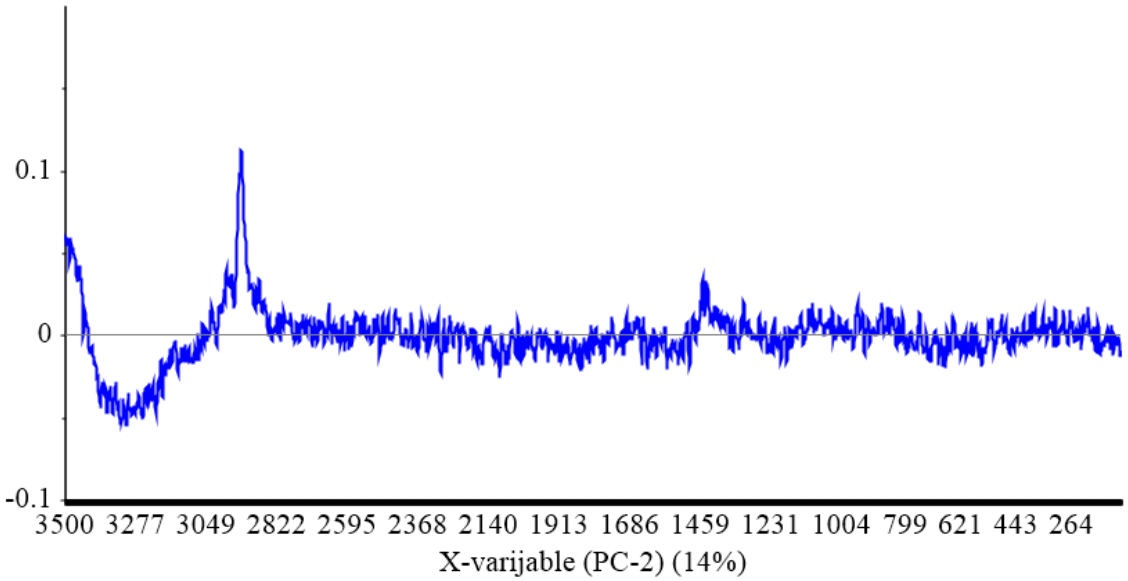
Slika 193. Raspodjela faktorskih bodova PC1 i PC2 matematički obrađenih (SNV i uklanjanje trenda polinomom 4. stupnja) Raman spektara PMPS C u području $\tilde{\nu}$ = 3500 - 100 cm^{-1} sa faktorskim opterećenjem za PC1 i PC2.

Dvodimenzionalni prikaz faktorskih bodova i opterećenja (Slika 193.) je dobar način za analizu odnosa između varijabli i identifikaciju najutjecajnijih varijabli u formiranju PCA modela. Međutim, u ovom slučaju obzirom na lošu mogućnost interpretacije, koristila su se linijska

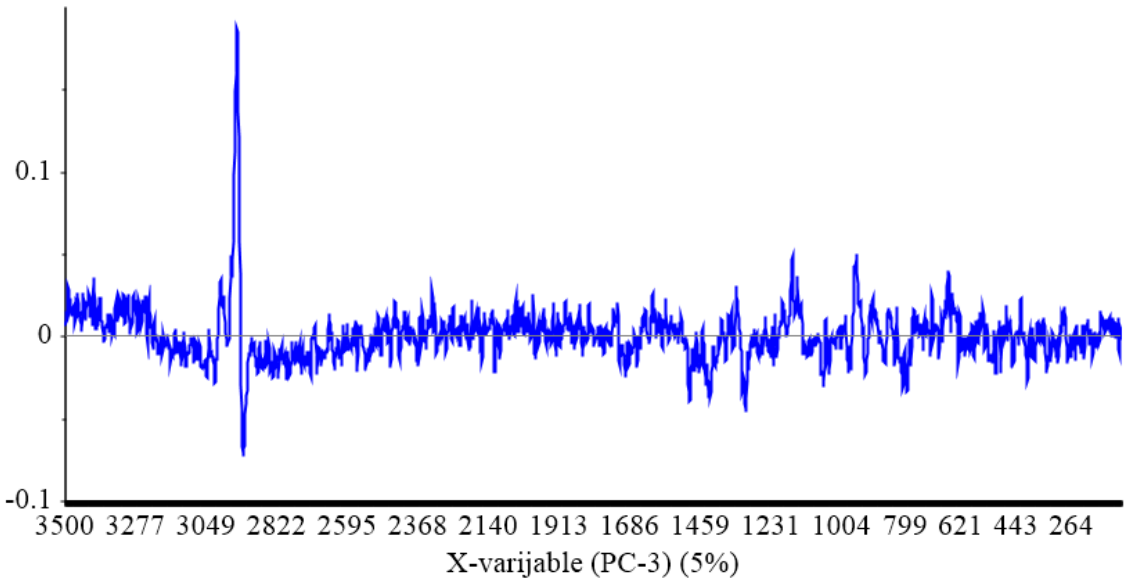
opterećenja koja imaju profil sličan originalnim spektralnim podacima, pa omogućuju odličnu interpretaciju opterećenja.



Slika 194. PC1 opterećenja po valnim brojevima ($\tilde{\nu}$) dobivena PCA analizom Raman spektara PMPS C.



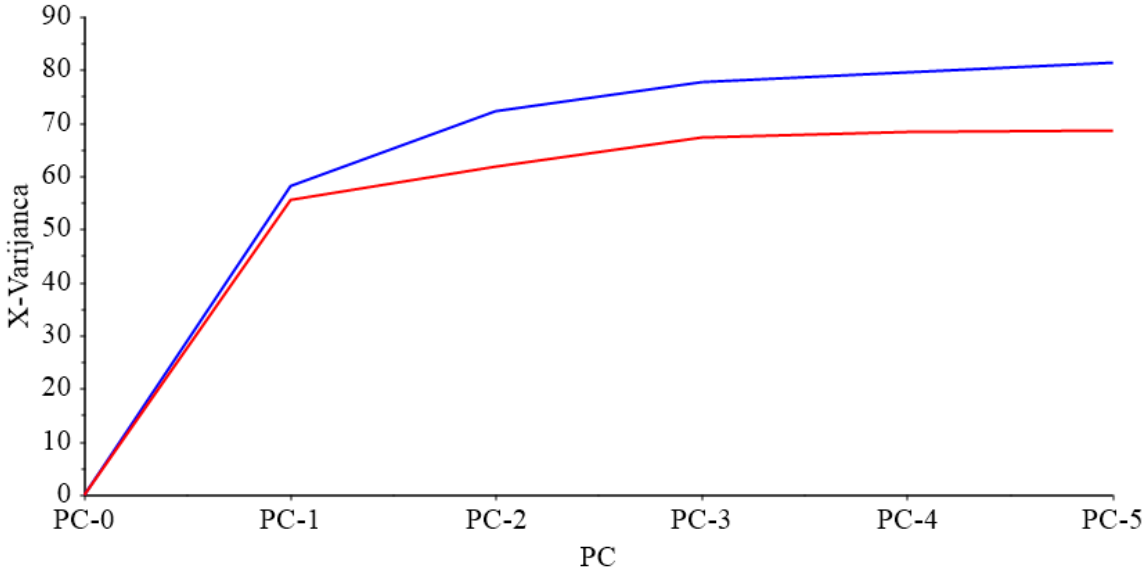
Slika 195. PC2 opterećenja po valnim brojevima ($\tilde{\nu}$) dobivena PCA analizom Raman spektara PMPS C.



Slika 196. PC3 opterećenja po valnim brojevima ($\tilde{\nu}$) dobivena PCA analizom Raman spektara PMPS C.

Na slikama 194. - 196. mogu se identificirati varijable tj. najvažnije Raman spektralne regije, koje imaju najveći utjecaj na PC1, PC2 i PC3.
Na Slikama 136. - 138. može se vidjeti da valni brojevi koji su imali najveći utjecaj na formiranje PMPS C modela pripadaju spektralnim vrpcama oko $\tilde{\nu}$ = 2930 cm^{-1} proizlaze od CH_2 asimetričnog istezanja; $\tilde{\nu}$ = 1500 - 1200 cm^{-1} C-H/CH_2 deformacijske vibracije; $\tilde{\nu}$ = 1150 - 1070 cm^{-1} proizlazi od C-O-C asimetričng istezanja; $\tilde{\nu}$ = 800 - 100 cm^{-1} deformacijske vibracije C-C-O grupa (Larkin, 2018)..

Kumulativnom kalibracijskom i validacijskom varijancom određen je optimalan broj PC-ova tj. dimenzionalnost modela, (Slika 197.), što je ključan preduvjet za formiranje kvalitetnog i robustnog PCA modela.



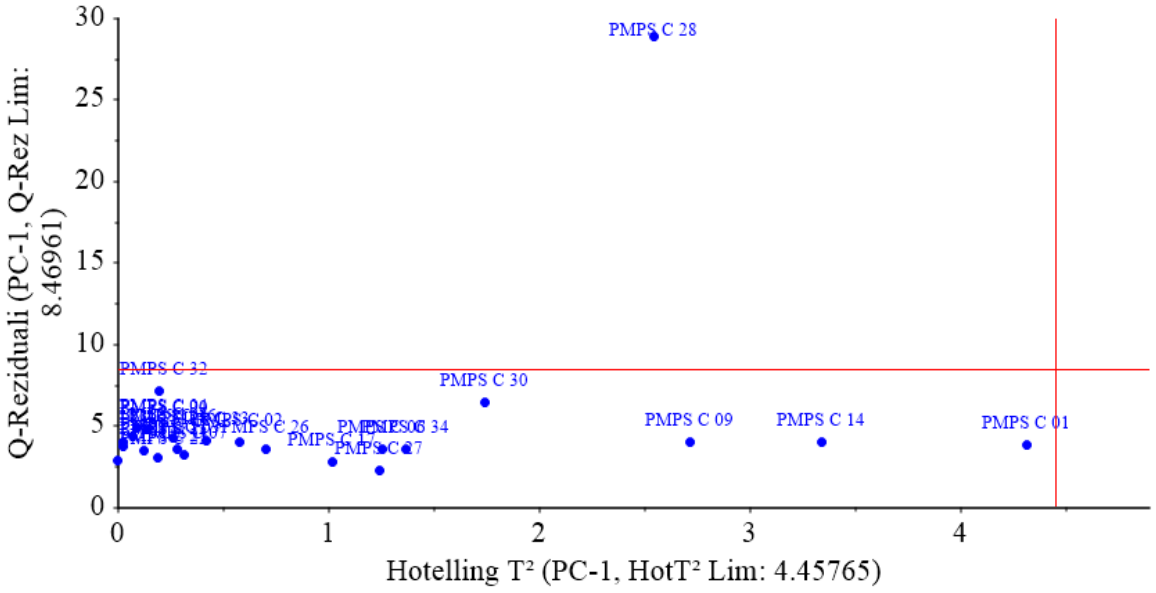
Slika 197. Kumulativna kalibracijska (plava) i validacijska (crvena) varijanca za svaki PC.

Slika 197. prikazuje koliko varijance opisuju različite glavne komponente (PC). Tri glavne komponente (PC1 i PC2, PC3) opisuju ukupno 78 % ukupne kalibracijske varijance , te 67% validacijske varijance. Dobivena razlika među kalibracijskom i validacijskom varijancom može ukazivati na prisutnost netipičnih uzoraka u kalibracijskom skupu podataka. Potrebno je načiniti dodatne statističke analize kako bi se detaljnije analizirali uzorci kalibracijskog skupa PMPS C.

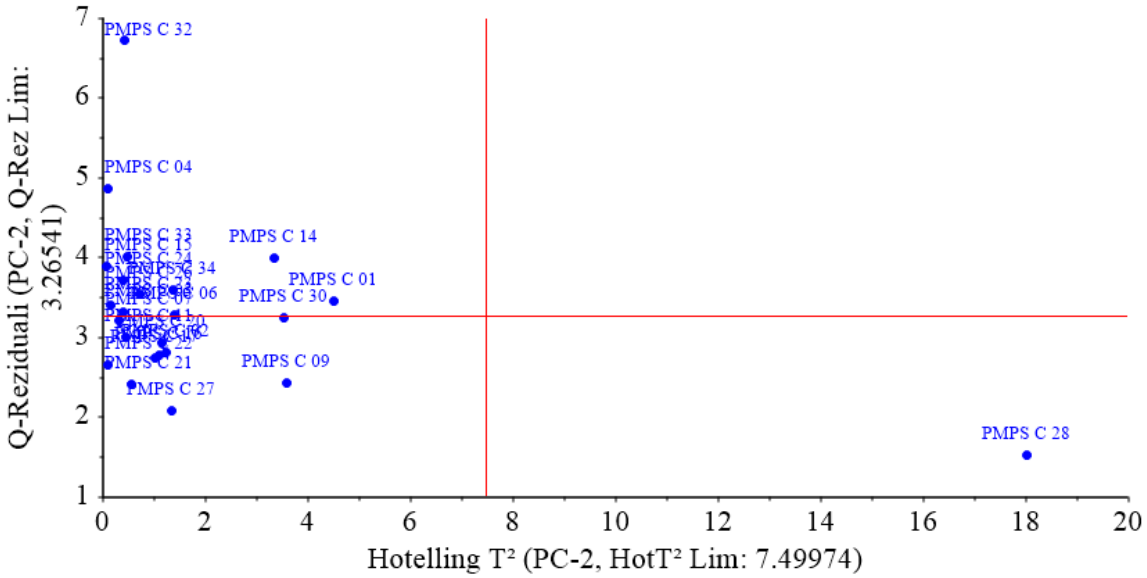
Tablica 11. Kumulativna kalibracijska i validacijska varijanca za svaki PC.

	PC0	PC1	PC2	PC3	PC4	PC5
Kalibracija	0	58.2561	72.3655	77.6649	79.6894	81.4187
Validacija	0	55.6503	61.9239	67.3490	68.3851	68.4992

Pomoću Hotelling T^2 statistike i Q-reziduala identificirani su netipični uzorci unutar kalibracijskog seta uzoraka PMPS C (ovdje ispod).



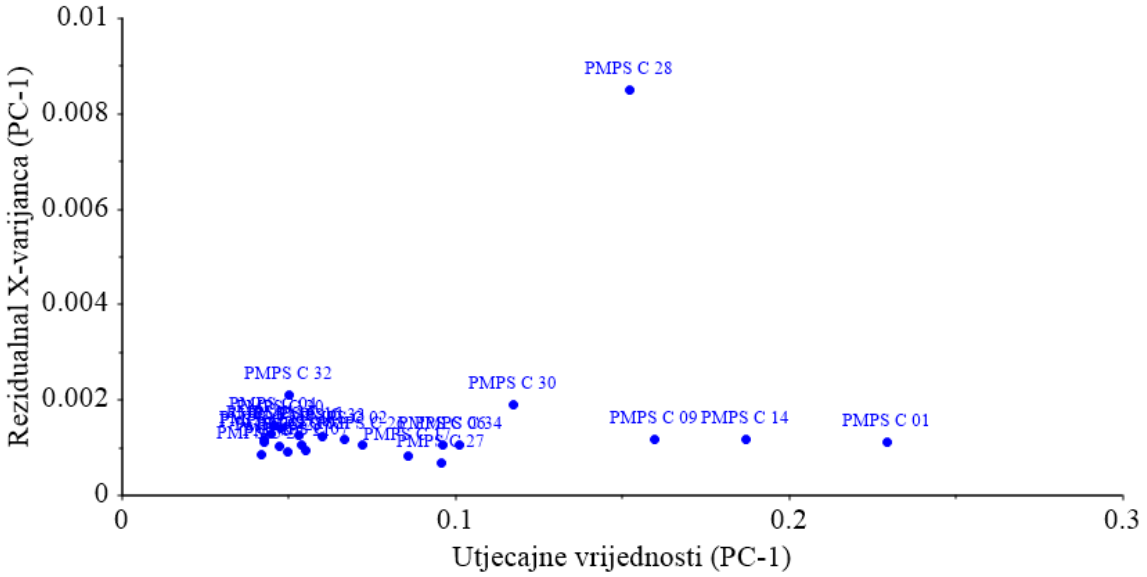
Slika 198. Hotelling T^2 statistika i Q-reziduali uzoraka PMPS C za PC1. Crvene linije predstavljaju kritične granice sa razinom značajnosti od 5 %.



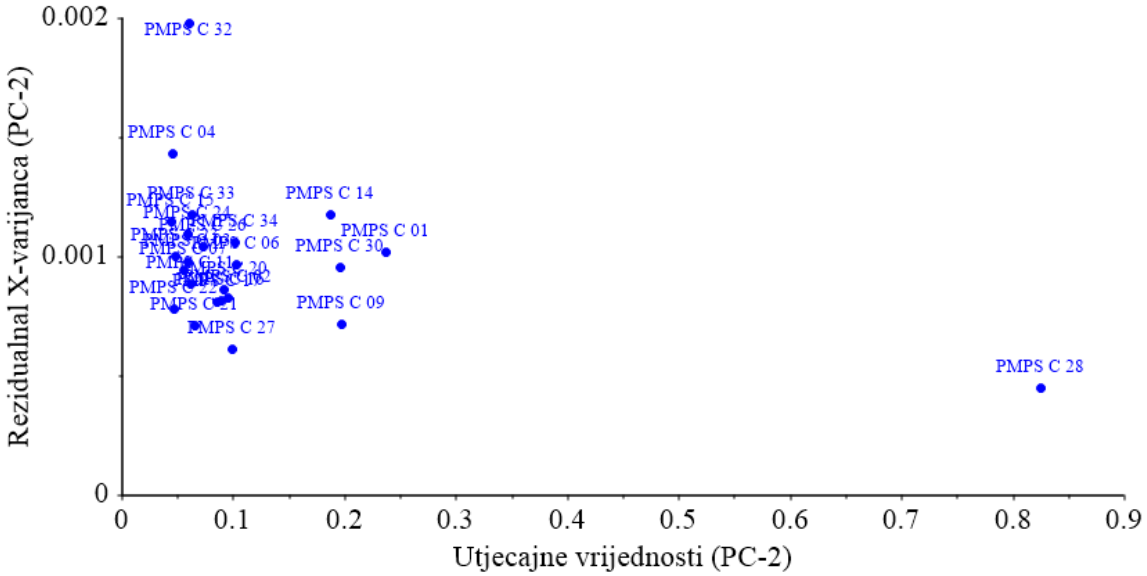
Slika 199. Hotelling T^2 statistika i Q-reziduali uzoraka PMPS C za PC2. Crvene linije predstavljaju kritične granice sa razinom značajnosti od 5 %.

Q i Hotelling T^2 statistika koriste se u svrhu karakterizacije varijabilnosti unutar trening seta uzoraka. Na Slikama 198. i 199. jasno se vidi da uzorak PMPS C 28 ima visoku Hotelling T^2 vrijednost i visoke Q reziduale i može se smatrati netipičnim uzorkom. Međutim, bilo je

potrebno dodatno istražiti Hotelling T^2 vrijednost i Q-reziduale svih PMPS C uzoraka iz ovoga seta.



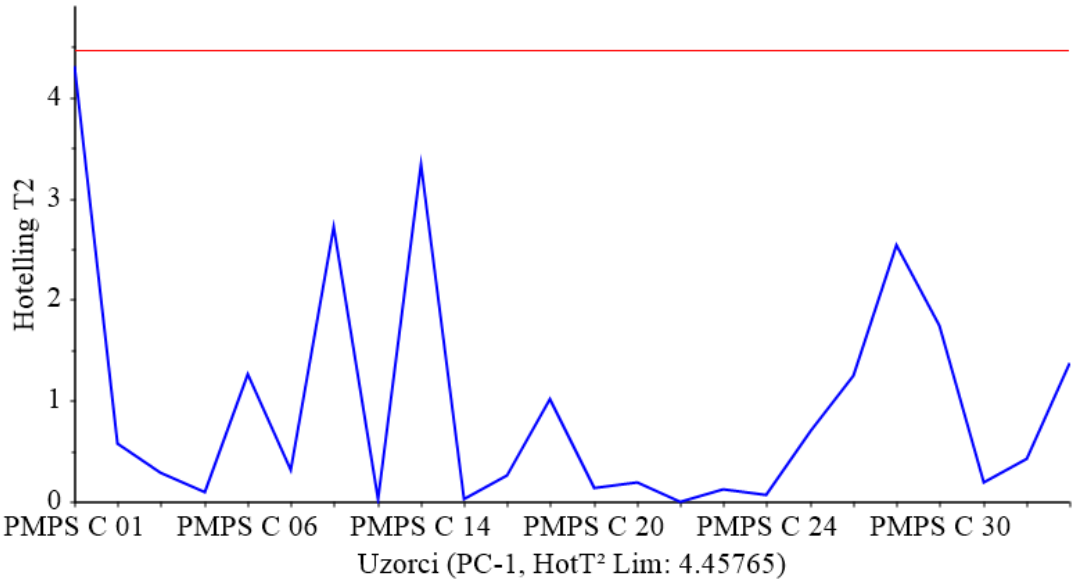
Slika 200. Rezidualna X-varianca i utjecajna vrijednost uzoraka PMPS C za PC1.



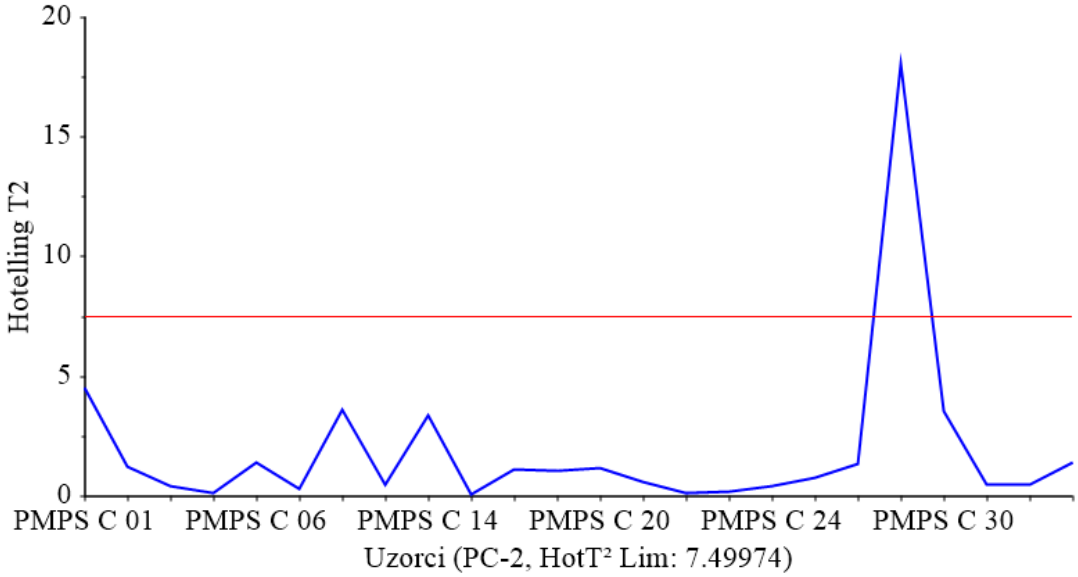
Slika 201. Rezidualna X-varianca i utjecajna vrijednost uzoraka PMPS C za PC2.

Na Slikama 200. i 201. prikazana je rezidualna X-varijanca uzoraka PMPS C sa utjecajem svakog pojedinog uzorka na PCA model. Jasno se može vidjeti da uzorak PMPS C 28 ima visoku rezidualnu X-varijancu. Ovaj uzorak ima i visok utjecaj na drugu glavnu komponentu

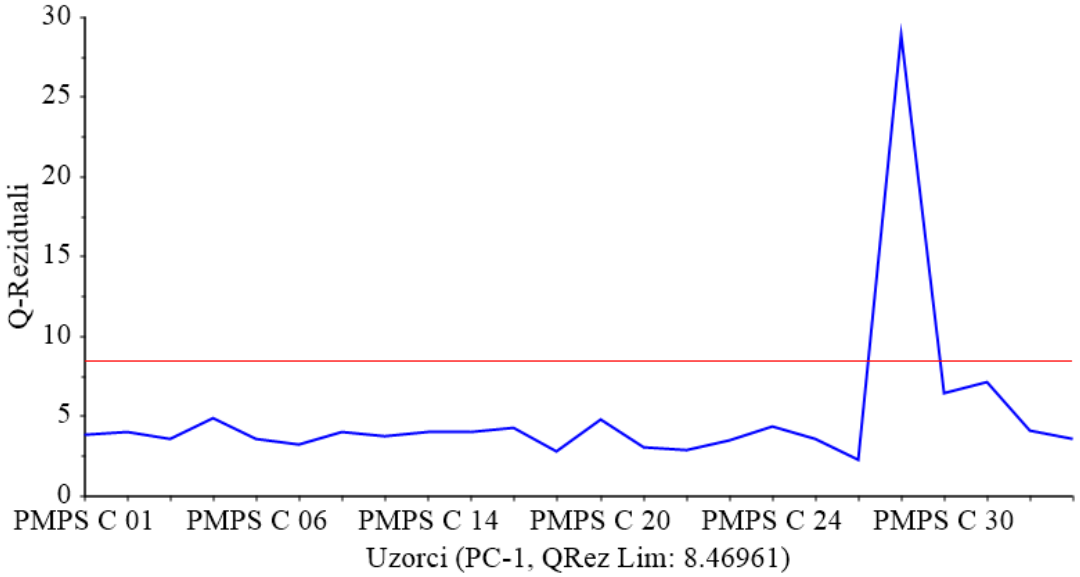
(PC2) PCA modela, što znatno utječe na formiranje robusnog i pouzdanog PCA modela, pa je i ovdje utvrđeno da je ovaj uzorak (PMPS C 28) potrebno nadalje razmotriti kao potencijalni netipični uzorak.



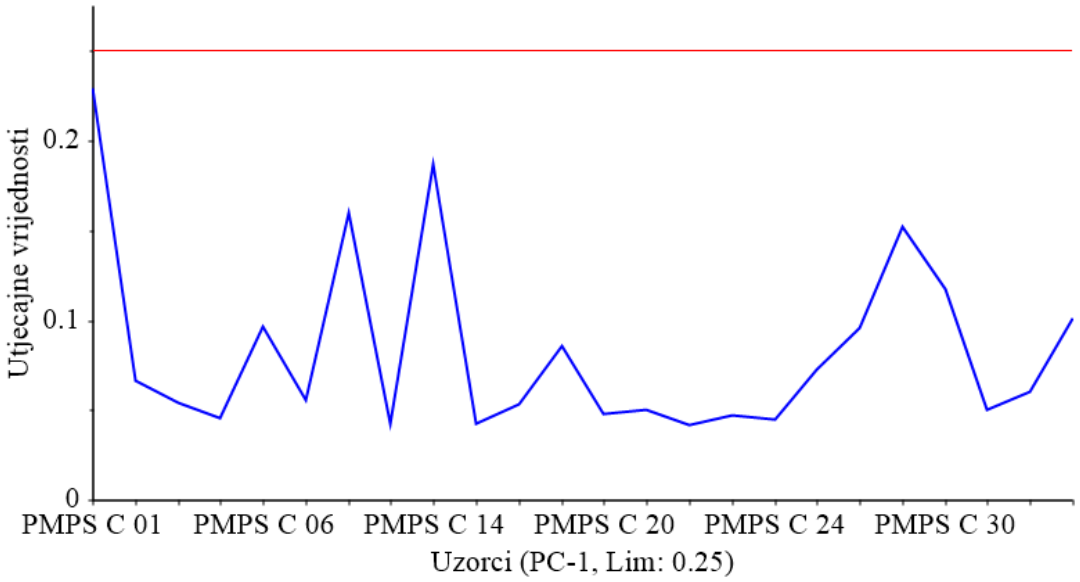
Slika 202. Hotelling T^2 statistika uzoraka PMPS C za PC1 sa pripadajućom kritičnom vrijednosti (crvena linija).



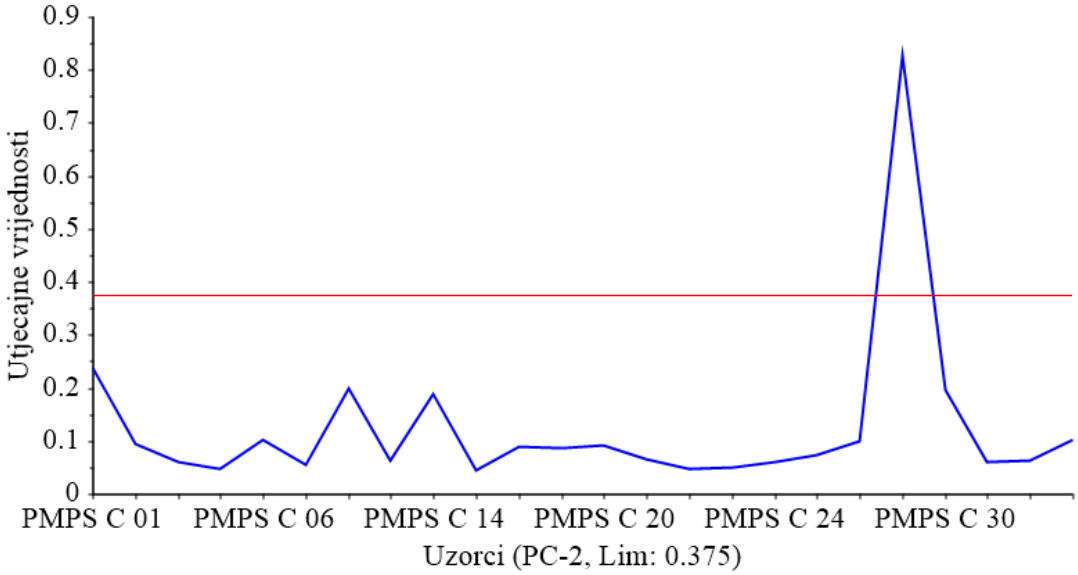
Slika 203. Hotelling T^2 statistika uzoraka PMPS C za PC2 sa pripadajućom kritičnom vrijednosti (crvena linija).



Slika 204. Q-reziduali uzoraka PMPS C za PC 1 s pripadajućom graničnom linijom (crvena linija).



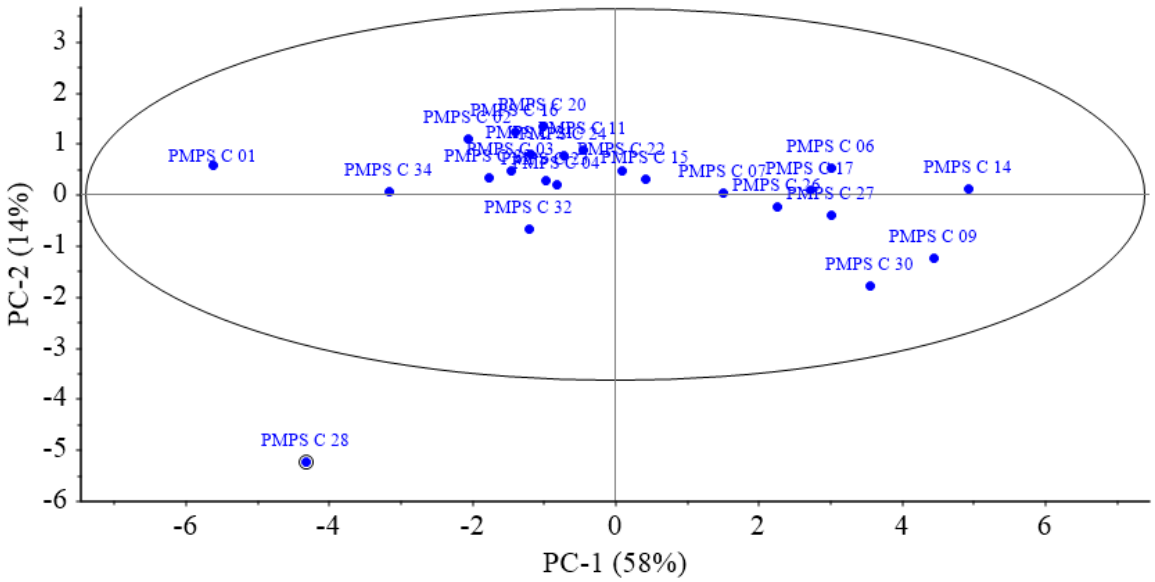
Slika 205. Utjecajne vrijednosti uzoraka PMPS C za PC1 sa pripadajućom kritičnom vrijednosti (crvena linija).



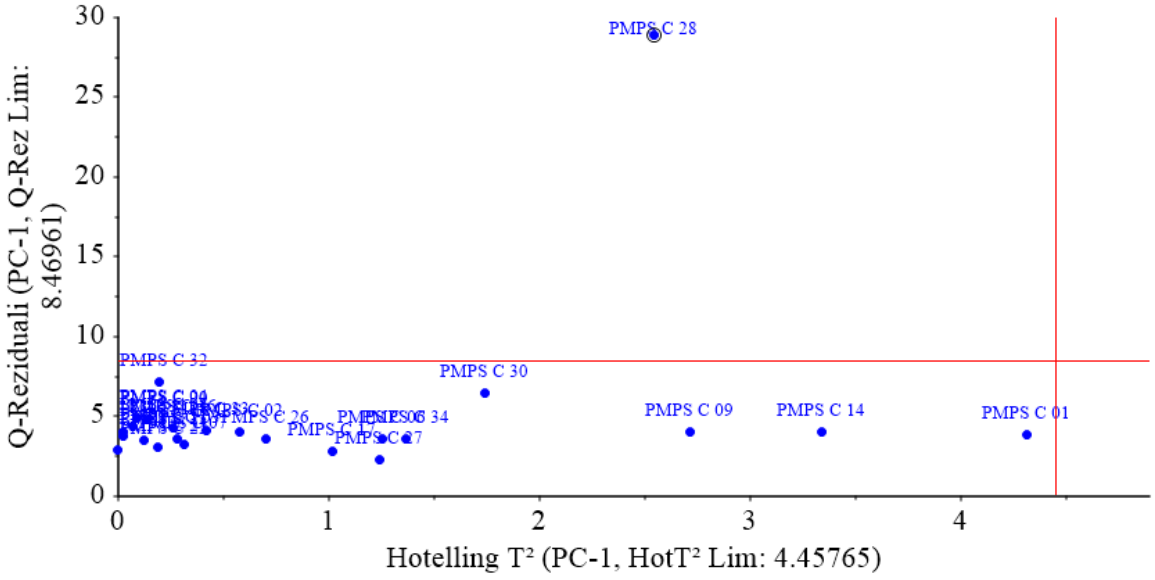
Slika 206. Utjecajne vrijednosti uzoraka PMPS C za PC2 sa pripadajućom kritičnom vrijednosti (crvena linija).

Na Slikama 202. - 206. jasno se može vidjeti da uzorak PMPS C 28 ima visoke visoke vrijednosti Q reziduala na PC1, te visoke vrijednosti Hotelling T^2 na PC2 i također visok utjecaj na PC2.
Dakle, uzorak PMPS C 28 nedvojbeno je netipični uzorak te je izuzet iz skupa uzoraka PMPS C za kalibraciju i optimizaciju PCA modela.

Kako je statističkom analizom Raman spektralnih podataka za PMPS C utvrđeno da svakako treba izdvojiti uzorak PMPS C 28 kao netipični uzorak, bilo je potrebno nanovo načiniti PCA model preostalih Raman spektralnih podataka PMPS C i to bez ovoga netipičnog uzorka.

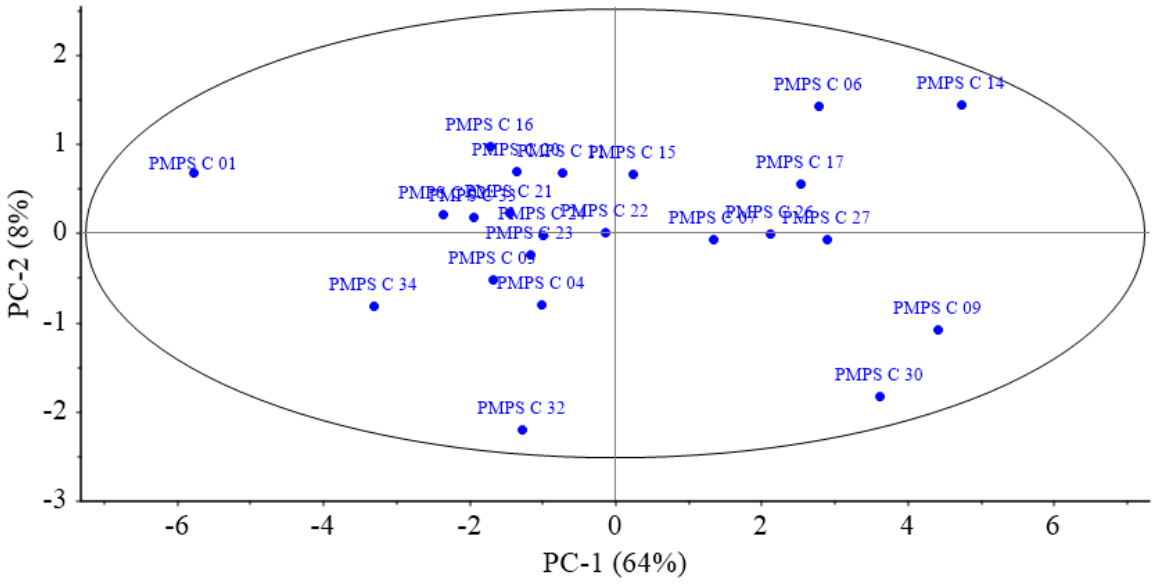


Slika 207. Raspodjela faktorskih bodova PC1 i P 2 matematički obrađenih (SNV i uklanjanje trenda polinomom 4. stupnja) Raman spektara PMPS C u području $\tilde{\nu}$ = 3500 - 100 cm^{-1} sa Hotelling T^2 elipsom (interval pouzdanosti 95 %) i označenim netipičnim uzorkom.



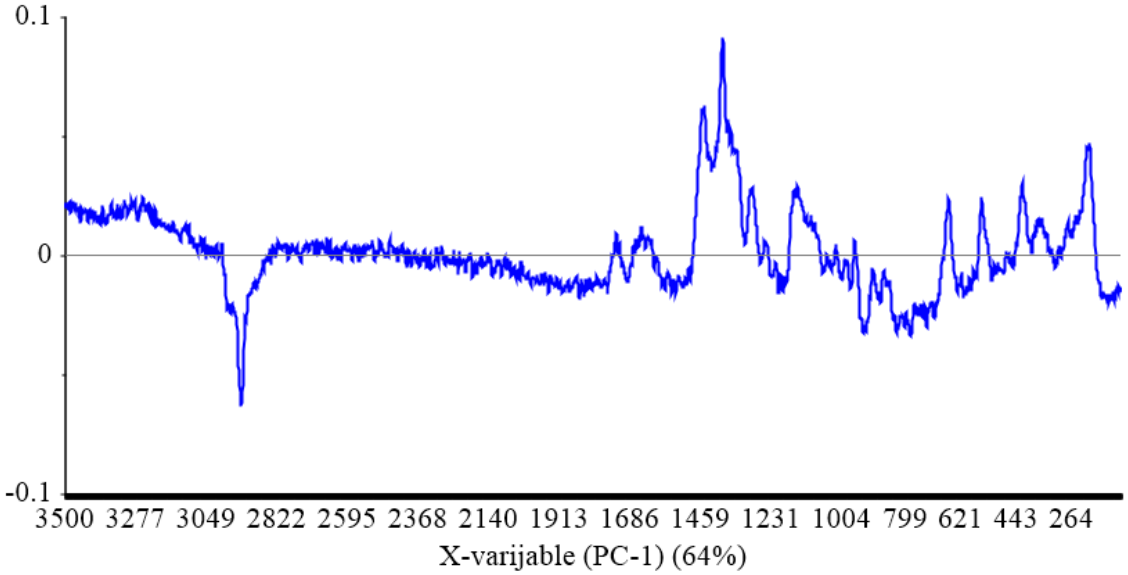
Slika 208. Hotelling T^2 statistika i Q-reziduali uzoraka PMPS C za PC1 s označenim netipičnim uzorkom i pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija).

Nakon uklanjanja uzorka PMPS C 28, koji je identificiran kao netipični uzorak, bilo je potrebno ponoviti fromiranje PCA modela s preostalim Raman spektralnim podacima.

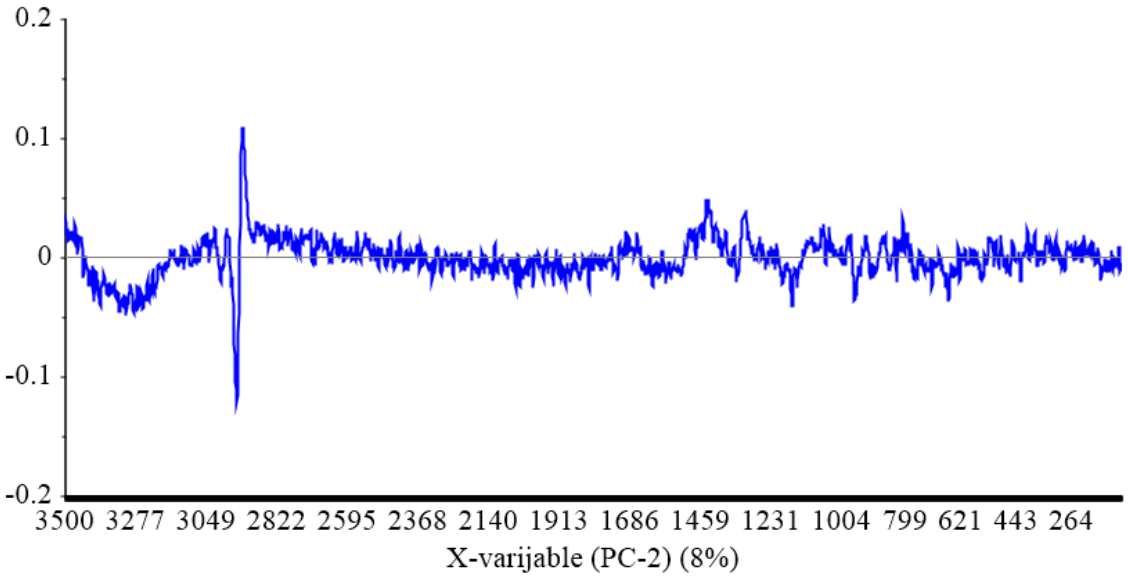


Slika 209. Raspodjela faktorskih bodova PC1 i PC2 matematički obrađenih (SNV i uklanjanje trenda polinomom 4. stupnja) Raman spektara PMPS C u području $\tilde{\nu}$ = 3500 - 100 cm^{-1} sa Hotelling T^2 elipsom (interval pouzdanosti 95 %) nakon uklanjanja netipičnog uzorka.

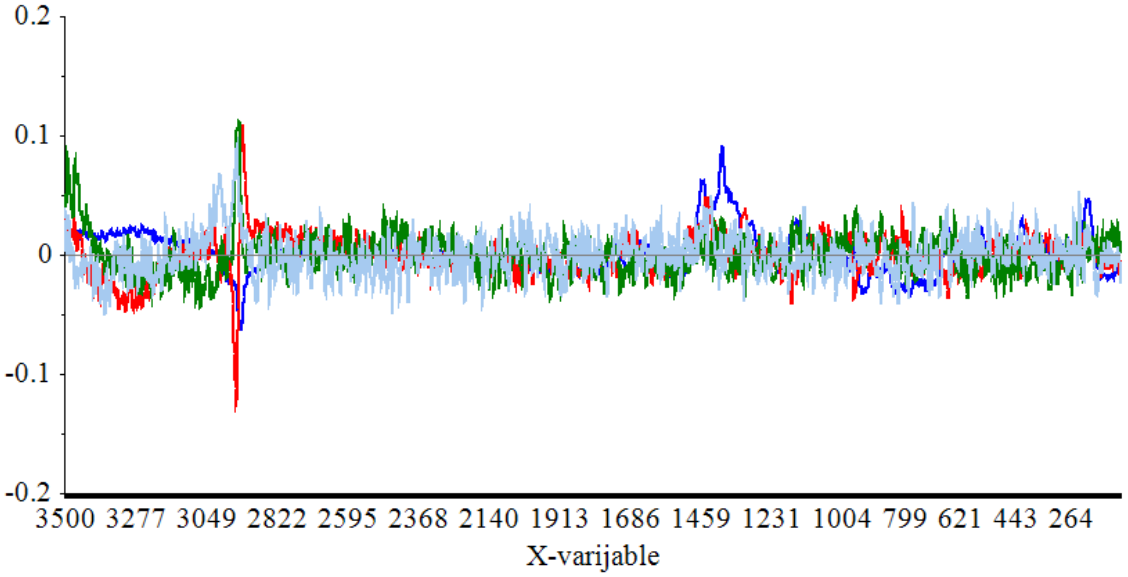
Iz raspodjele faktorskih bodova (Slika 209.) jasno se vidi jednolika raspodjela PMPS C uzoraka kroz cijelo područje. Nema prisutnih trendova među uzorcima niti se uzorci međusobno nejednoliko grupiraju. Svi PMPS C uzorci unutar su intervala pouzdanosti 95 % te nema uzoraka koji predstavljeju potencijalne netipične odnosno ekstremne PMPS C uzorke.
Nakon uklanjanja netipičnog uzorka (PMPS C 28) načinjena su i linijska opterećenja za PC1, PC2 te su prikazani valni brojevi ($\tilde{\nu}$) s najvećim opterećenjima koji najviše doprinose pojedinoj komponenti (PC1 i PC2) (ovdje ispod).



Slika 210. PC1 opterećenja za po valnim brojevima ($\tilde{\nu}$) dobivena PCA analizom Raman spektara PMPS C.



Slika 211. PC2 opterećenja za po valnim brojevima ($\tilde{\nu}$) dobivena PCA analizom Raman spektara PMPS C.



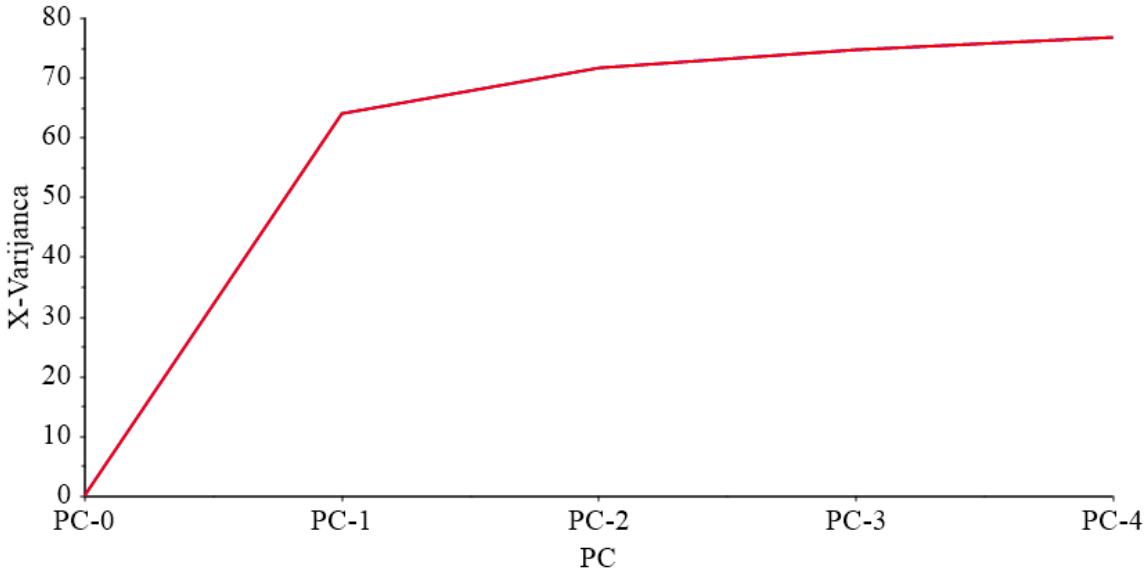
Slika 212. PC1 (plavo), PC2 (crveno), PC3 (zeleno) i PC4 (sivo) opterećenja po valnim brojevima dobivena PCA analizom za skup od 34 NIR spektara PMPS C.

Preklopljena opterećenja za PC1, PC2, PC3 i PC4 zorno ukazuju na karakteristične spektralne regije, koje definiraju glavne komponente (Slika 212.).

Iz slike 212. se može vidjeti da PC1 i PC2 obuhvaćaju većinu važnih Raman spektralnih podataka, dok druge dvije glavne komponente (PC3 i PC4) obuhvaćaju uglavnom šum i nepotrebne su za formiranje ovoga PCA modela.

Na Slikama 210. - 212. može se vidjeti da najveći doprinos u formiranju PMPS C modela pripada spektralnim vrpcama oko $\tilde{\nu}$ = 2930 cm^{-1} koje proizlaze od CH_2 asimetričnog istezanja; $\tilde{\nu}$ = 1500 - 1200 cm^{-1} proizlaze od C-H/CH_2 deformacijske vibracije; $\tilde{\nu}$ = 1150 - 1070 cm^{-1} proizlazi od C-O-C asimetričng istezanja; $\tilde{\nu}$ = 800 - 100 cm^{-1} od deformacijske vibracije C-C-O grupa (Larkin, 2018).

Kumulativnom kalibracijskom i validacijskom varijancom određen je optimalan broj PC-ova (Slika 213.), što je ključan preduvjet za daljnje formiranje modela.



Slika 213. Kumulativna kalibracijska (plava) i validacijska (crvena) varijanca za svaki PC.

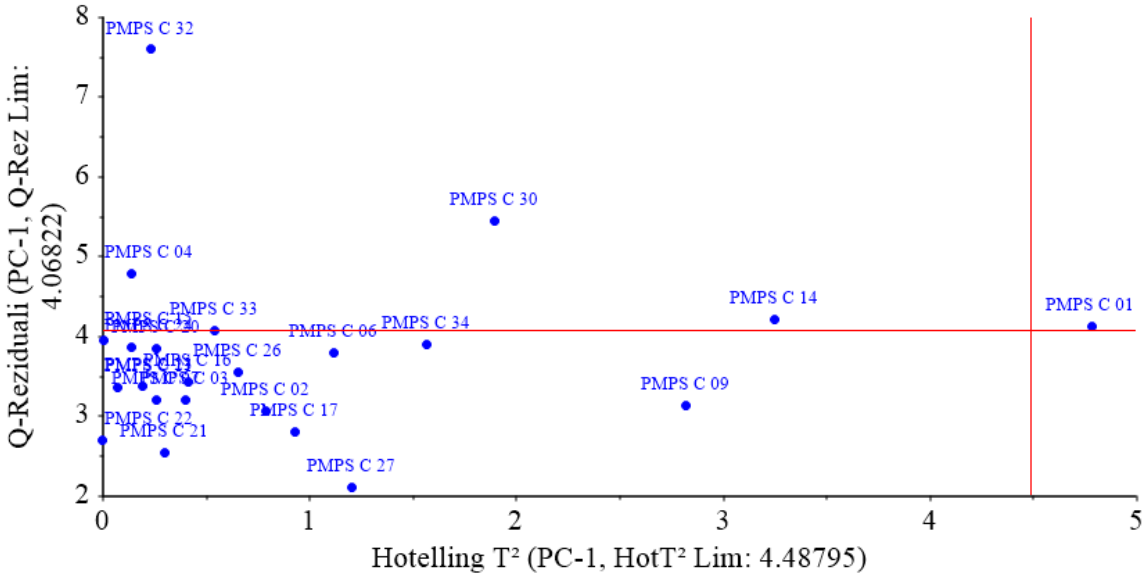
Tablica 12. Kumulativna kalibracijska i validacijska varijanca za svaki PC nakon uklanjanja netipičnog uzorka.

	PC 0	PC1	PC2	PC3	PC4
Kalibracija	0	63,9365	71,6610	74,7069	76,8302
Validacija	0	63,9259	71,6443	74,6846	76,8029

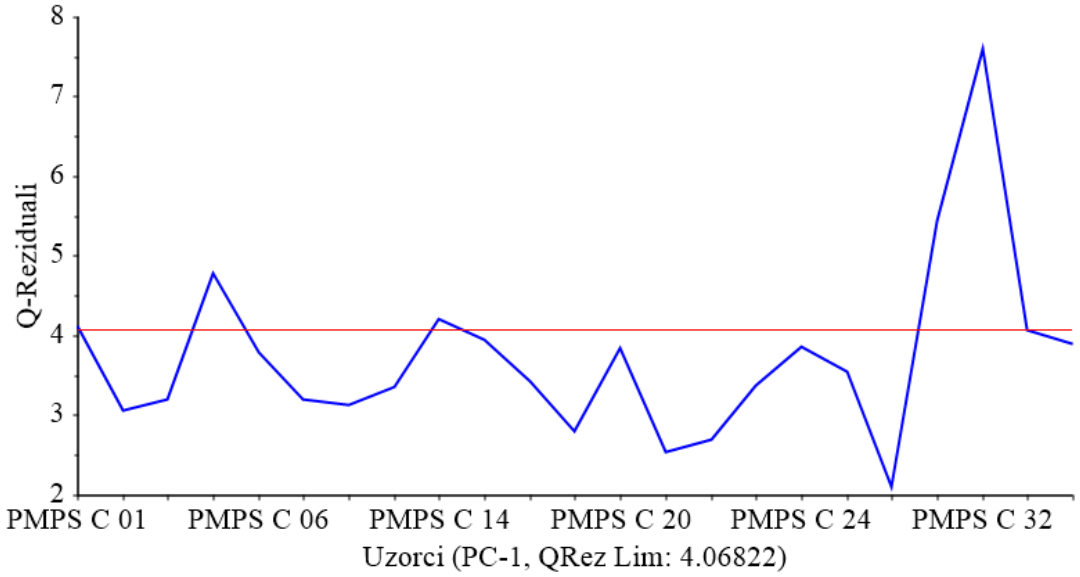
Slika 213. prikazuje koliko varijance opisuju glavne komponente (PC). Odabir broja glavnih komponenti ima veliki utjecaj na model - ukoliko se odabere pre mali broj PC-a, smanjuje se specifičnost modela, dok preveliki broj PC-a vodi ka smanjenoj osjetljivosti PCA modela. Kumulativna kalibracijska i validacijska varijanca se u potpunosti preklapaju (Slika 213.) što ukazuje na reprezentativne kalibracijske i validacijske skupove uzoraka. Također, ovaj rezultat upućuje i na nepostojanje netipičnih uzoraka unutar ovoga kalibracijskog seta uzoraka. Dvije glavne komponente (PC1 i PC2) opisuju ukupno 72 % ukupne varijance, dok tri glavne komponente (PC1, PC2 i PC3) opisuju ukupno 75 % ukupne varijance.

Važno je napomenuti da postotak objašnjene X.varijance nije krucijalan u procjeni kvalitete modela. Krajnji cilj modela je međusobno razdvajanje klasa, te ako su relevantne informacije za razdvajanje klasa sadržane u nekoliko prvih latentnih varijabli, objašnjena X varijanca, kod ovako kompleksnih skupova podataka koji uključuje veliki broj varijabli, može biti biti mala.

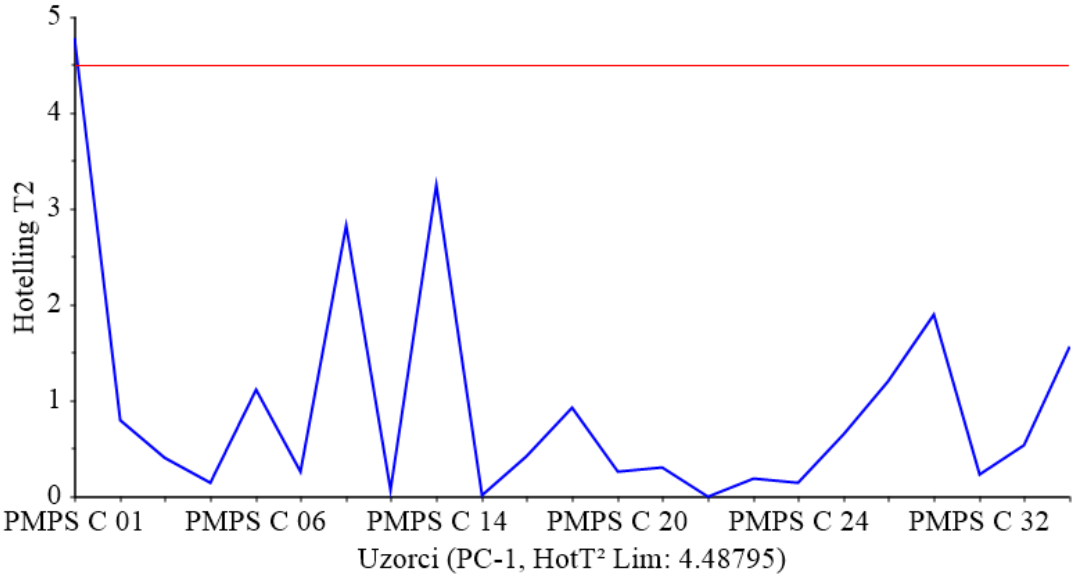
Prisutnost netipičnih uzoraka u modelu je procijenjena pomoću utjecajnih vrijednosti, Hotelling T^2 statistike i Q reziduala (ovdje ispod). Kod PCA, obje statistike (T^2 i Q) su dobro aproksimirane hi - kvadrat distribucijom.



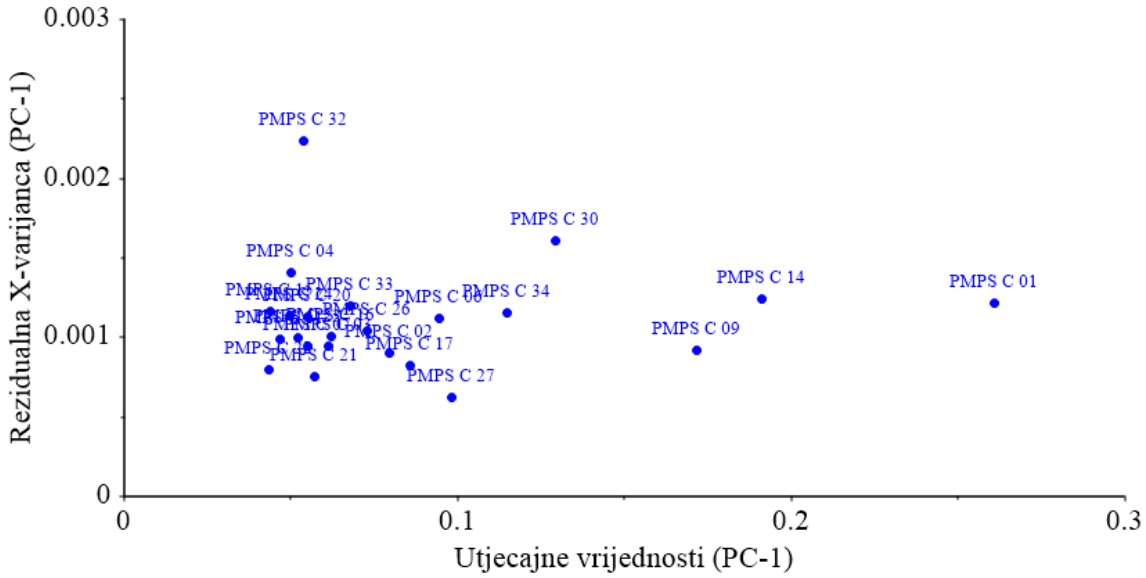
Slika 214. Hotelling T^2 statistika i Q-rezidual uzoraka PMPS C za PC1 sa pripadajućom kritičnom vrijednosti (crvena linija).



Slika 215. Q-reziduali uzoraka PMPS C za PC1 s pripadajućom graničnom linijom (crvena linija).



Slika 216. Hotelling T^2 statistika uzoraka PMPS C za PC1 sa pripadajućom kritičnom vrijednosti (crvena linija).

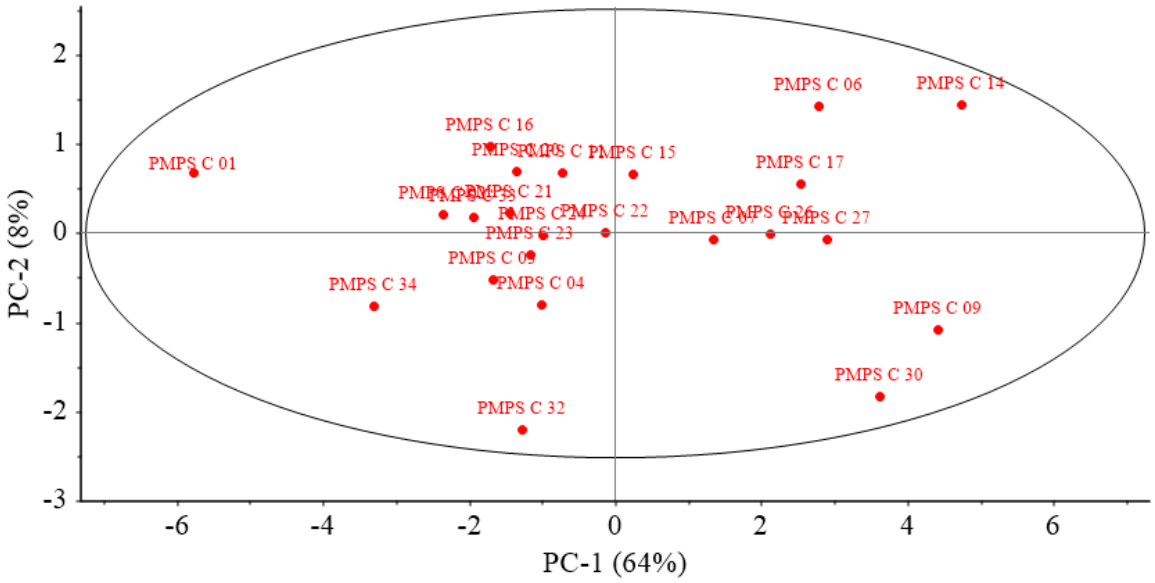


Slika 217. Rezidualna X-varianca i utjecajna vrijednost uzoraka PMPS C za PC1.

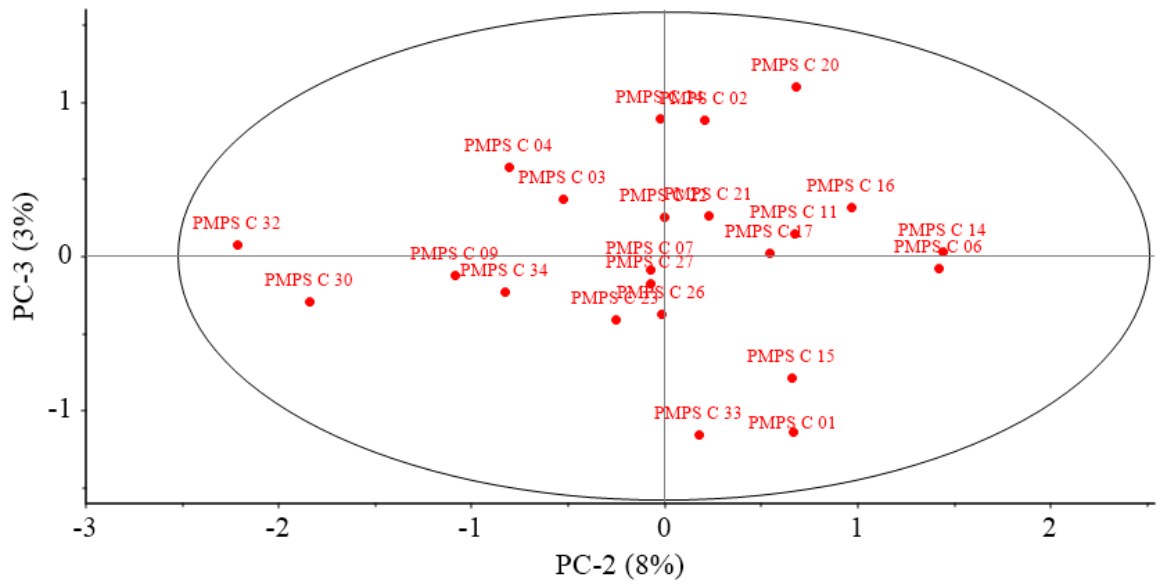
Iz dobivenih rezultata prikazanih na Slikama 214. - 217. može se vidjeti da u ovome setu nema prisutnih PMPS C uzoraka koje bi trebalo dodatno analizirati kao potencijalne netipične uzorke.

4.3.5.3 Optimizacija Raman SIMCA modela

Optimizacija SIMCA modela provedena je zasebno za svaku serogrupu polisaharida (PMPS A i PMPS C) i to uz korištenje trening setova uzoraka *Venetian blind* postupkom unakrsne validacije. Postupak unakrsne validacije proveden je sa sedam validacijskih grupa, kako je to opisano u poglavlju 2.5.3.2. Na temelju optimizacije Raman SIMCA modela odabran je broj od tri PC-a za svaki pojedinačni model - PMPS A i PMPS C.



Slika 218. Raspodjela faktorskih bodova PC2 i PC3 validacijskih grupa matematički obrađenih (SNV i uklanjanje trenda polinomom 4. stupnja) Raman spektara PMPS C u području $\tilde{\nu}$ = 3500 - 100 cm^{-1} sa sa Hotelling T^2 elipsom (interval pouzdanosti 95 %) nakon provedene unakrsne validacije.



Slika 219. Raspodjela faktorskih bodova PC2 i PC3 validacijskih grupa matematički obrađenih (SNV i uklanjanje trenda polinomom 4. stupnja) Raman spektara PMPS C u području $\tilde{\nu}$ = 3500 - 100 cm^{-1} sa sa Hotelling T^2 elipsom (interval pouzdanosti 95 %) nakon provedene unakrsne validacije.

Na Slikama 218. – 219. koji prikazuju raspodjelu faktorskih bodova validacijskih PMPS C uzoraka, jasno se vidi jednolika raspodjela kroz cijelo područje, te skoro potpuno poklapanje sa kalibracijskim uzorcima. Nema prisutnih trendova među uzorcima niti se uzorci međusobno nejednoliko grupiraju. Svi PMPS C uzorci unutar su intervala pouzdanosti 95 %.

4.3.5.4 Validacija Raman SIMCA modela

Zbog relativno malog broja uzoraka PMPS A i C nije bilo moguće posebno odvojiti uzorke za kalibraciju, optimizaciju i evaluaciju već je provedena unakrsna validacija i na temelju ove validacije je odabran optimalan broj PC faktora. Za procjenu klasifikacijske sposobnosti Raman SIMCA modela provedena je validacija vanjskim setom uzoraka PMPS A i C na modelima formiranim uz pomoć trening seta ovih uzoraka i optimiranim unakrsnom validacijom. Vanjski set uzoraka PMPS A i C nije sudjelovao niti u kalibraciji niti u optimizaciji ovoga modela.
Vanjski set uzoraka se sastojao od devetnaest spektara PMPS A i C i još dodatno po pet spektara PMPS Y i W135, koji su bili negativne probe. Rezultati dobiveni za SIMCA model sa tri PC-a za svaku klasu (serogrupu) su prikazani u Tablici 13. kao matrica zabune. Za procjenu

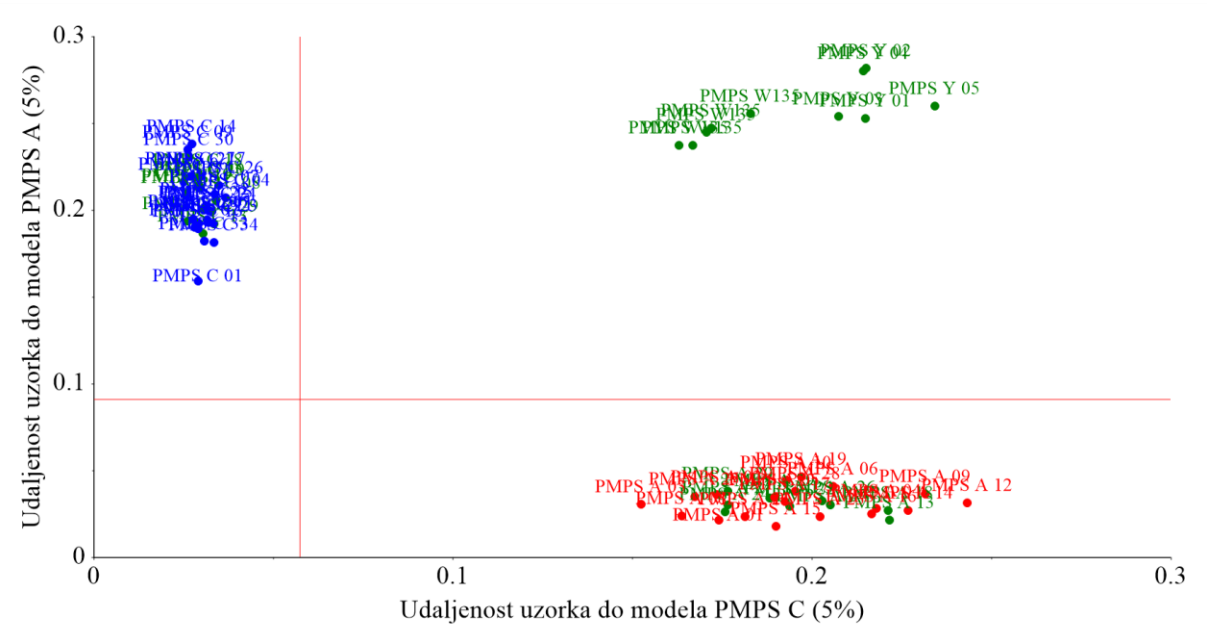
prediktivne sposobnosti Raman SIMCA klasifikacijskog modela razmatrani su validacijski parametri, osjetljivost, specifičnost, i učinkovitost za svaku ciljnu klasu PMPS A i C kao i za ukupi Raman SIMCA model. Rezultati dobiveni za negativne probe W135 i Y korišteni su za demonstraciju specifičnosti formiranoga modela.

U tablici 13. prikazani su rezultati klasifikacije Raman SIMCA klasifikacijskog modela. Iz tablice se može jasno vidjeti da su svi uzorci vanjskog seta ispravno klasificirani.

Tablica 13. Matrica zabune za Raman SIMCA model sa tri PC-a po klasi za vanjski test set uzoraka PMPS A i PMPS C.

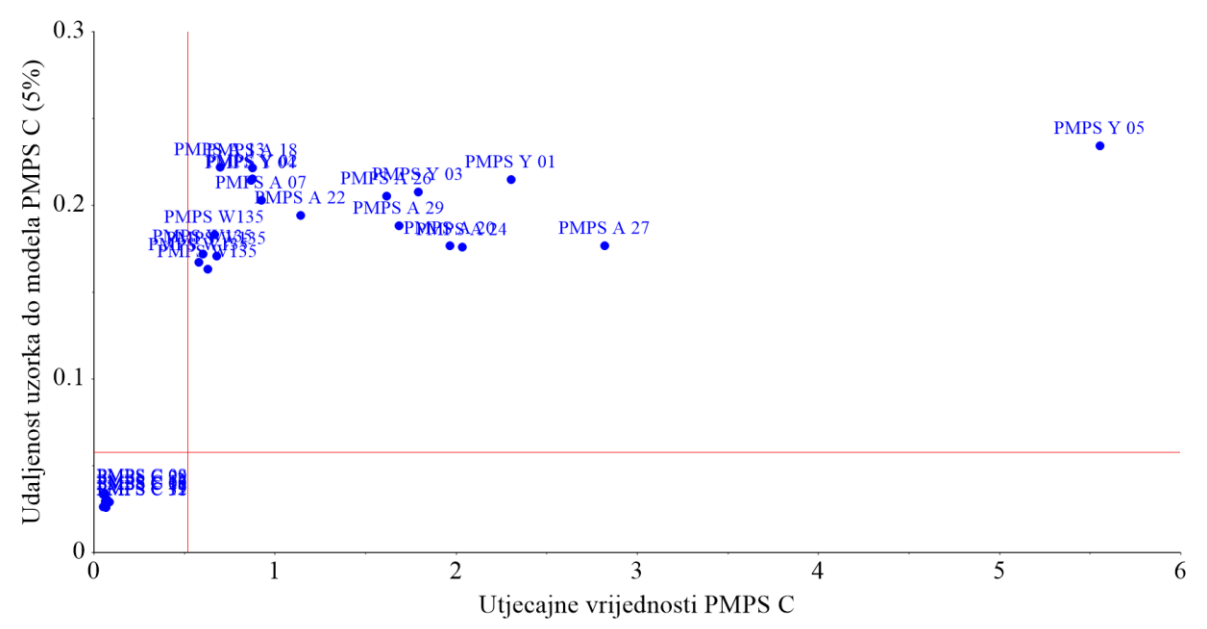
stvarno/predviđeno	klasa PMPS A	klasa PMPS C	nije klasificirano
klasa PMPS A	9	0	0
klasa PMPS C	0	10	0
klasa PMPS W135	0	0	5
klasa PMPS Y	0	0	5
CSNS	100%	100%	TSNS = 100%
CSPS	100%	100%	TSPS = 100%
CEFF	100%	100%	TEFF (3 PC) = 100%

CSNS, CSPS, CEFF, TSNS, TSPS i TEFF su opisani u poglavlju 2.5.3.2.

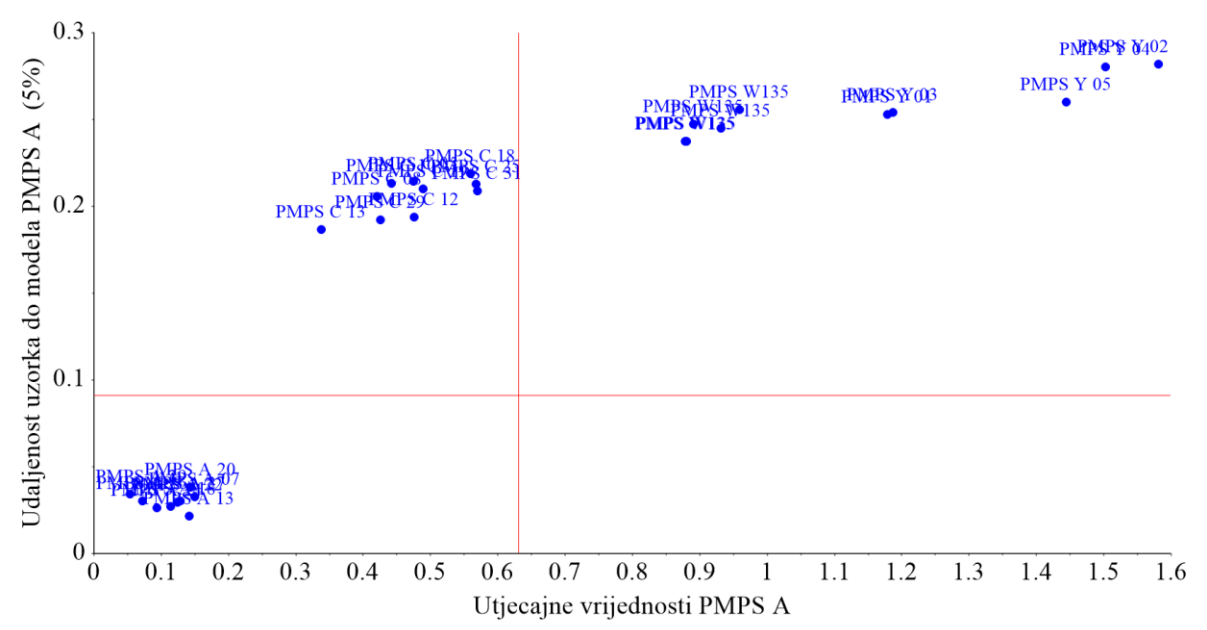


Slika 220. Cooman dijagram s pripadajućim graničnim vrijednostima (okomita i vodoravna crvena linija)

Slika 220. prikazuje Cooman dijagram za Raman SIMCA klasifikacijski model temeljen na prethodno definiranim PCA modelima za PMPS A i PMPS C (nivo pouzdanosti 5 %). Na Slici 220. može se vidjeti da su svi uzorci nedvosmisleno identificirani i klasificirani u odgovarajuće klase PMPS A i C te da nema uzoraka između dvije klase PMPS. Uzorci negativnih proba PMPS W135 i Y nisu klasificirani niti u jedu klasu PMPS. Udaljenost između PCA modela svake klase su relativno velike, što ukazuje na to da je dobra razlučivost ovih klasa.



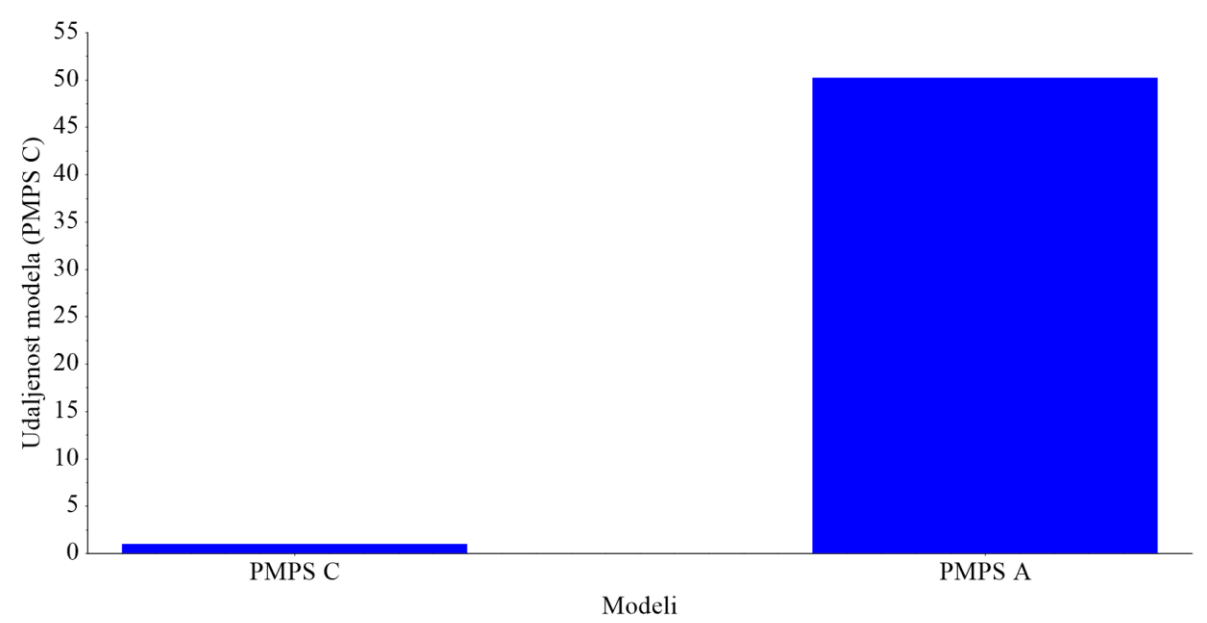
Slika 221. Udaljenosti uzorka od modela PMPS C (Si) s utjecajnim vrijednostima ovih uzoraka (Hi) za model PMPS C.



Slika 222. Udaljenosti uzorka od modela PMPS A (Si) s utjecajnim vrijednostima ovih uzoraka (Hi) za model PMPS A.

Nepoznati uzorci su uspoređeni s PCA modelom i to uz korištenje dvaju parametara - udaljenosti uzorka do PCA modela (Si) i utjecajnosti uzoraka (Hi). Hi vrijednost opisuje koliko bi uzorak imao utjecaja na model kada bi bio uključen u model.

Slike 221. - 222. ukazuju na udaljenost uzoraka meningokoknih polisaharida PMPS A, C, W135 i Y od PCA modela u Raman SIMCA modelu. Uzorci svake klase (serogrupe) iz kalibracijskog seta udaljeni su od PCA modela druge klase i potpuno se razlikuju jedan od drugog. Niti jedan uzorak nije kategoriziran pogrešno u drugu klasu ili istovremeno u obje klase (serogrupe). Zbog toga se može zaključiti da je SIMCA klasifikacija bazirana na PCA modeliranju izuzetno učinkovita u klasifikaciji PMPS A i C.

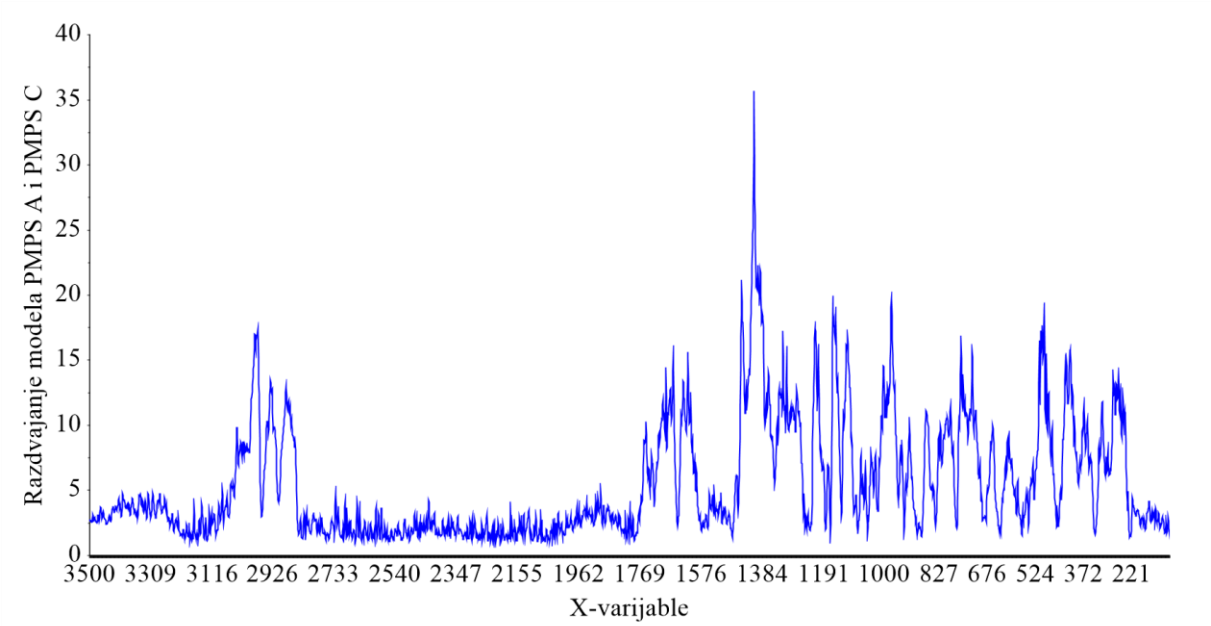


Slika 223. Udaljenost dvaju Raman SIMCA modela.

Tablica 14. Udaljenost Raman SIMCA modeal

	PMPS A	PMPS C
PMPS A	1	50.2282
PMPS C	50.2282	1

Na Slici 223. i u Tablici 14. prikazana je udaljenost između pojedinih PCA modela i ova slika i tablica zorno pokazuju da je udaljenost između ova dva modela velika i da su modeli jasno odvojeni.



Slika 224. Diskriminacijska moć različitih valnih brojeva ($\tilde{\nu}$) Raman spektara u diskriminaciji PMPS A i PMPS C.

Na slici 224. prikazana je diskriminacijska moć različitih valnih brojeva ($\tilde{\nu}$) u diskriminaciji dviju klasa PMPS A i C. Valni brojevi koji su imali najveći utjecaj na diskriminaciju klasa pripadaju spektralnim regijama oko $\tilde{\nu}$ = 2980 cm^{-1} koje proizlaze od C-H i CH_2 istezanja; $\tilde{\nu}$ = 2930 cm^{-1} proizlaze od CH_2 asimetričnog istezanja; $\tilde{\nu}$ = 2875 cm^{-1} proizlaze od CH_2 simetričnog istezanja $\tilde{\nu}$ = 1750 - 1700 cm^{-1} C=O vibracija istezanja karbonilne grupe te od $\tilde{\nu}$ = 1680 - 1639 cm^{-1} proizlaze od C=ONHR istezanja; $\tilde{\nu}$ = 1500 - 1200 cm^{-1} C-H/CH_2 deformacijske vibracije; $\tilde{\nu}$ = 1350 - 1140 cm^{-1} P=O istezanja; $\tilde{\nu}$ = 1150 - 1070 cm^{-1} proizlazi od C-O-C asimetričng istezanja; $\tilde{\nu}$ = 980 - 965 cm^{-1} proizlaze od C-H savijanja van ravnine; a $\tilde{\nu}$ = 800 – 100 cm^{-1} deformacijske vibracije C-C-O grupa (Larkin, 2018).

Na temelju rezultata u ovom poglavlju (poglavlje 4.3.5) može se zaključiti kako je primjena Raman spektroskopije u kombinaciji sa SIMCA modelom, koji se zasniva na formiranju pojedinačnih PCA modela za svaki PMPS (A i C) zasebno, visoko učinkovita za identifikaciju ovih pročišćenih meningokoknih polisaharida.

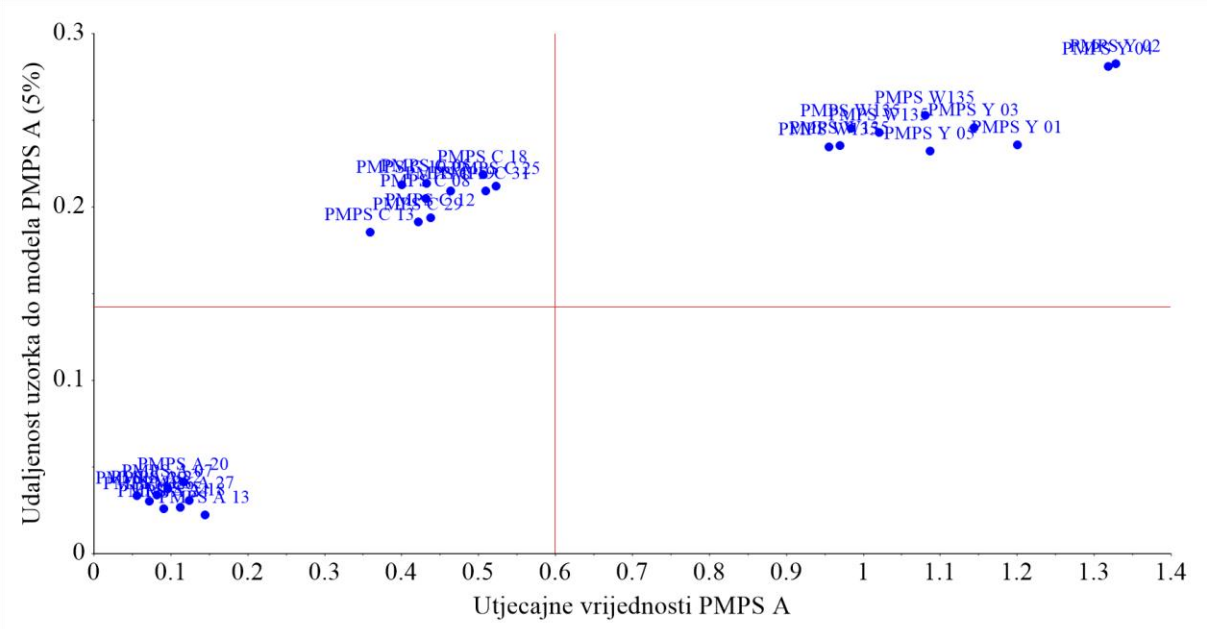
SIMCA model pripada skupini klasnog modeliranja, odnosno svaka serogrupa uzoraka meningokoknih polisaharida modelirana je zasebno reprezentativnim skupom proizvodnih uzoraka u pojedinačni PCA model. SIMCA model je izrazito osjetljiv na prisutstvo netipičnih uzoraka, pa su ovi uzorci identificirani primjenom različitih metoda statističke analize i izuzeti iz konačnog PCA modela. Na ovako formiranom PCA modelu za svaku serogrupu PMPS dodatno je istražena mogućnost jednoklasne klasifikacije PMPS A i PMPS C gdje je provjerena

sposobnost klasifikacije odnosno identifikacije uzoraka samo jedne klase (serogrupe) od interesa, dok su se svi ostali uzorci PMPS identificirali kao nepripadajući oodabranoj klasi. Opisani pristup primijenjen je na jednoklasnu identifikaciju uzoraka PMPS A i PMPS C.

4.3.6. Jednoklasna klasifikacija - model PMPS A

Za procjenu klasifikacijske sposobnosti PMPS A modela, koji je formiran na temelju trening seta PMPS A uzoraka i optimiran unakrsnom validacijom, provedena je validacija vanjskim setom uzoraka. Ovaj skup uzoraka PMPS A nije sudjelovao u kalibraciji niti u optimizaciji modela PMPS A. Vanjski set se sastoji od devet spektara proizvodnih uzoraka PMPS A, deset spektara proizvodnih uzoraka PMPS C i još od po pet spektara PMPS Y i W135, koji su upotrebljeni kao negativne probe.

Prikazana je (Slika 225.) sposobnost identifikacije uzoraka PMPS A iz vanjskog validacijskog seta.



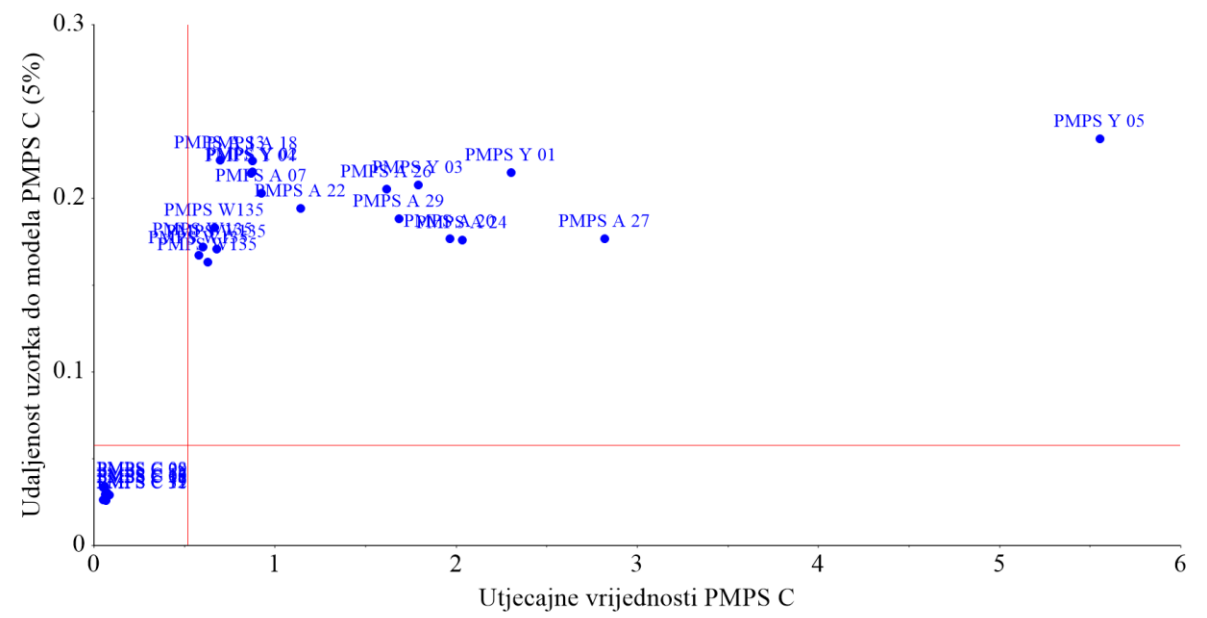
Slika 225. Udaljenosti uzorka PMPS A od modela (Si) s utjecajnim vrijednostima ovih uzoraka (Hi) za jednoklasni SIMCA model PMPS A.

Formirani jednoklasni SIMCA PMPS A model je uspješno identificirao sve uzorke iz vanjskog seta. Uzorci PMPS A identificirani su kao pripadnici ciljne klase, dok su svi preostali uzorci

PMPS C, W135, Y uspješno prepoznati kao nepripadajući uzorci. Niti jedan uzorak nije krivo identificiran i dodjeljen ciljnoj klasi.

4.3.7. Jednoklasna klasifikacija - model PMPS C

Za procjenu klasifikacijske sposobnosti jednoklasnog SIMCA PMPS C modela, koji je formiran sa trening setom PMPS C uzoraka i optimiran unakrsnom validacijom, provedena je validacija ovoga modela vanjskim setom uzoraka, koji je bio ekvivalentan setu kao i kod validacije PMPS A modela. Ovaj skup uzoraka nije sudjelovao u kalibraciji niti u optimizaciji ovog jednoklasnog SIMCA PMPS C modela.

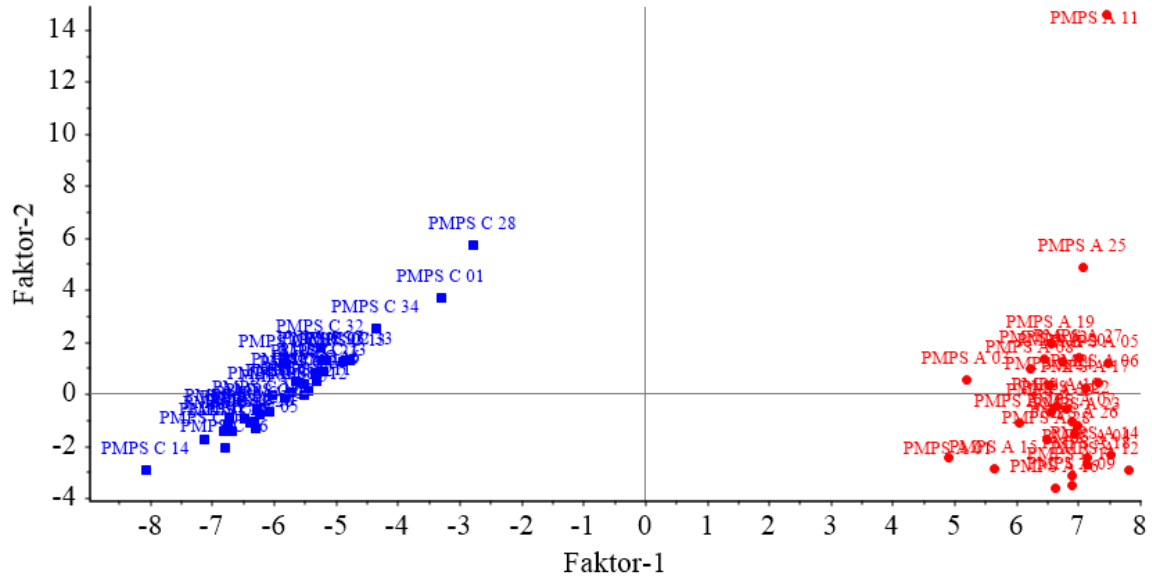


Slika 226. Udaljenosti uzorka PMPS C od modela (Si) s utjecajnim vrijednostima ovih uzoraka (Hi) za jednoklasni SIMCA model PMPS C.

Formirani PMPS C model je uspješno identificirao sve uzorke iz vanjskog seta. Uzorci PMPS C uspješno i nedvojbeno su identificirani kao pripadnici ciljne klase, dok su svi preostali uzorci PMPS A, W135, Y uspješno prepoznati kao nepripadajući uzorci. Niti jedan uzorak nije krivo identificiran i dodjeljen ciljnoj klasi

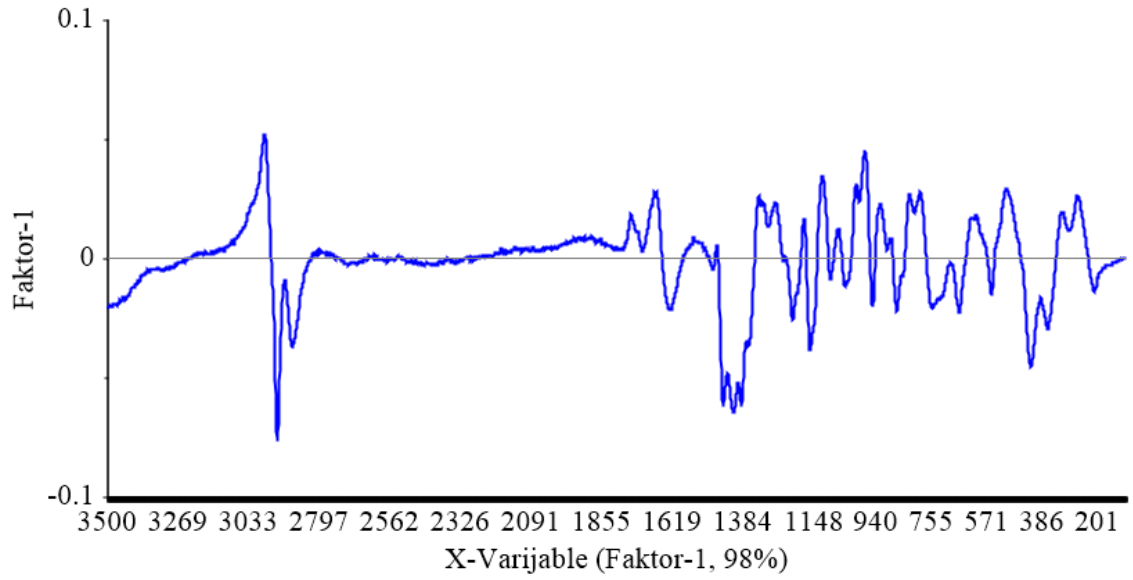
4.3.8. Raman PLS-DA model

PLS-DA klasifikacijski model kao orjentacijski prikaz koristi dijagram faktorskih bodova, na temelju kojeg se može dobiti uvid o mogućnosti razdvajanja serogrupa meningokoknih polisaharida ovim modelom. Obzirom da se PLS-DA model temelji na razlici među ovim serogrupama, prikaz faktorskih bodova bez pravilne validacije modela može dovesti do netočnih zaključaka o identifikacijskoj sposobnosti PLS-DA modela.



Slika 227. Raspodjela faktorskih bodova PC 1 i PC 2 matematički obrađenih (SNV i uklanjanje trenda polinomom 4. stupnja) Raman spektara PMPS A i C u području $\tilde{\nu}$ = 3500 - 100 cm^{-1}.

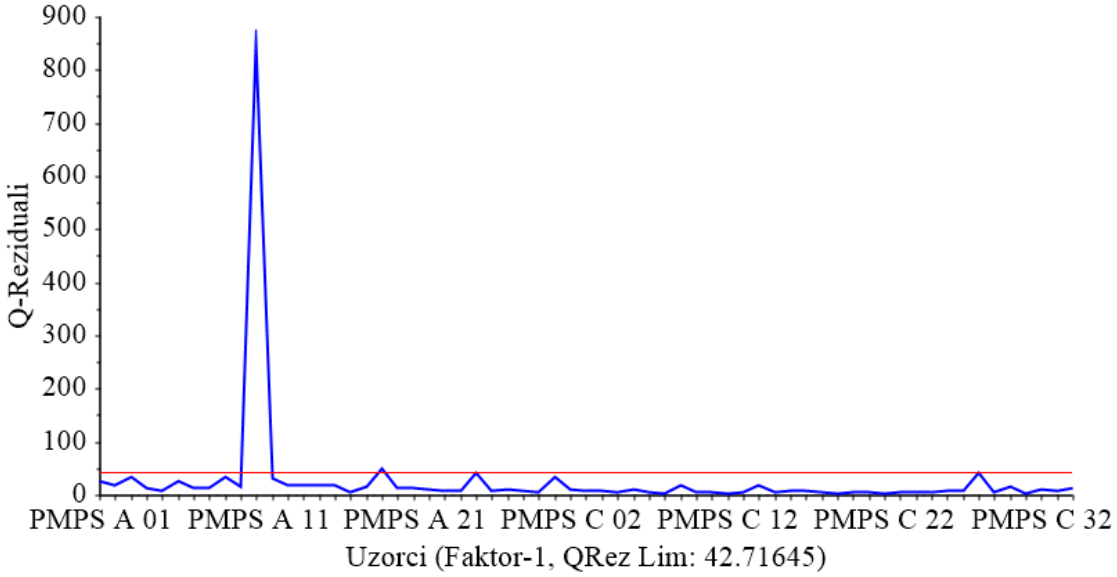
Na slici 227. se može vidjeti razdvajanje među PMPS A i C. Međutim, za izvođenje jasnog zaključka o ovome modelu, potrebna je daljnja analiza. Kako bi odredili najznačajnije valne brojeve ($\tilde{\nu}$), bilo je potrebno je prikazati dijagram opterećenja (Slika 228.).



Slika 228. Profil pterećenja za prvu latentnu varijablu (LV).

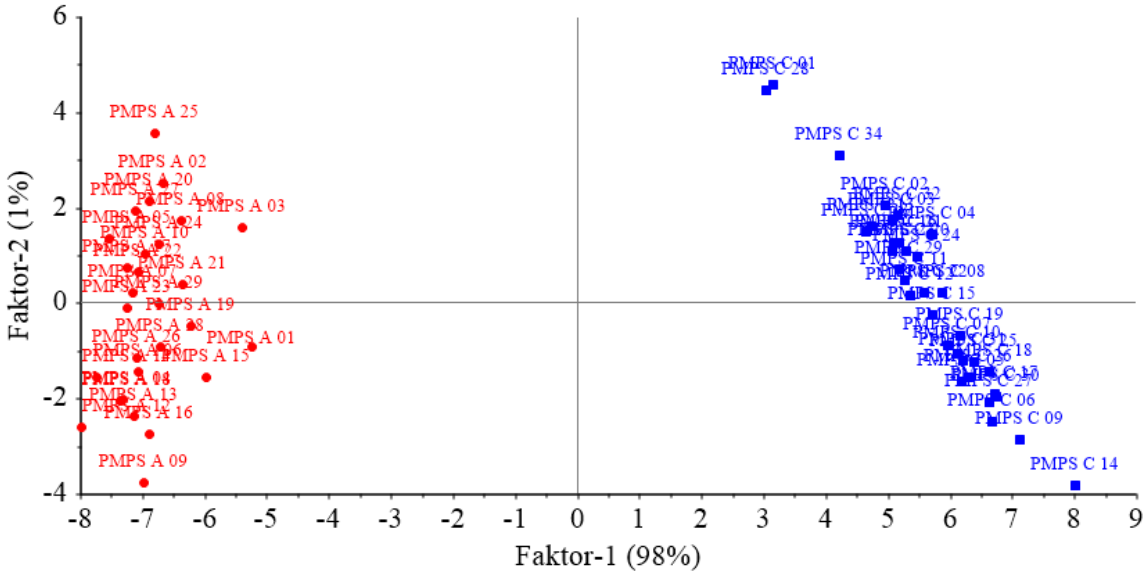
Na Slici 228. jasno su vidljive vrpce odgovorne za razdvajanje dviju klasa unutar modela, odnosno razdvajanje meningokoknih polisaharida serogrupe A i C. To su karakteristične spektralne vrpce pri: $\tilde{\nu}$ = 2980 cm^{-1} koje proizlaze od C-H i CH_2 istezanja; $\tilde{\nu}$ = 2930 cm^{-1} proizlaze od CH_2 asimetričnog istezanja; $\tilde{\nu}$ = 2875 cm^{-1} proizlaze od CH_2 simetričnog istezanja $\tilde{\nu}$ = 1750 - 1700 cm^{-1} C=O vibracija istezanja karbonilne grupe te od $\tilde{\nu}$ = 1680 - 1639 cm^{-1} proizlaze od C=ONHR istezanja; $\tilde{\nu}$ = 1500 - 1200 cm^{-1} C-H/CH_2 deformacijske vibracije; $\tilde{\nu}$ = 1350 - 1140 P=O istezanja; $\tilde{\nu}$ = 1150 - 1070 cm^{-1} proizlazi od C-O-C asimetričng istezanja; $\tilde{\nu}$ = 980 - 965 cm^{-1} proizlaze od C-H savijanja van ravnine; $\tilde{\nu}$ = 800 - 100 cm^{-1} deformacijske vibracije C-C-O grupa (Larkin, 2018).

Kako je PLS-DA model izuzetno osjetljiv na netipične uzorke, pomoću Q-reziduala statistički su identificirani netipični uzorci unutar kalibracijskog (trening) seta uzoraka PMPS A i C (Slika 229.)



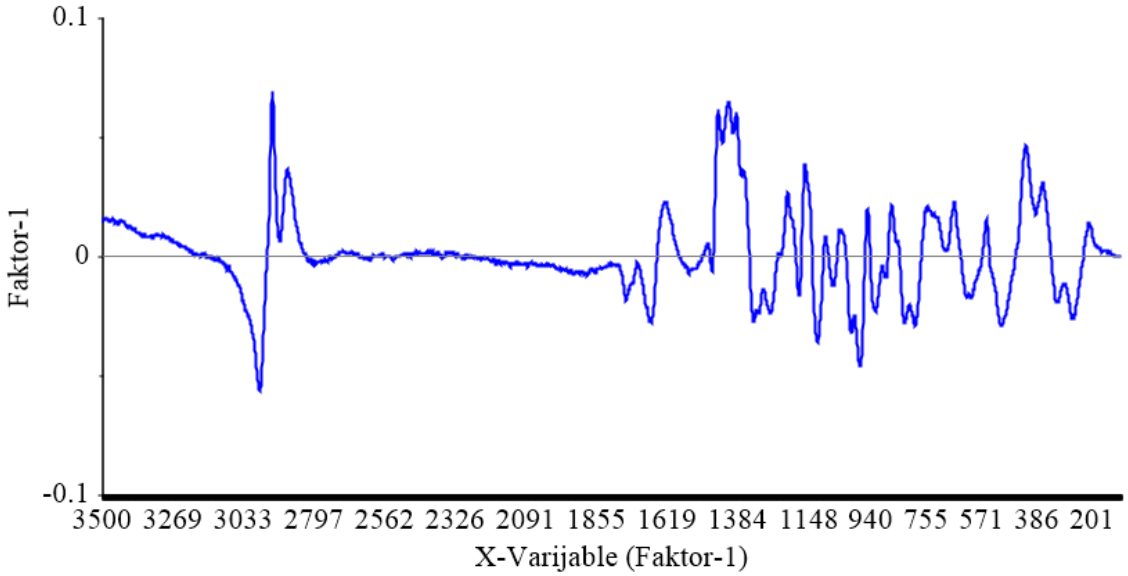
Slika 229. Q-reziduali uzoraka PMPS A i C s pripadajućom graničnom linijom (crvena linija). iz kalibracijskog seta uzoraka PMPS A i C.

Na slici 229. jasno se vidi da uzorak PMPS A 11 predstavlja netipični uzorak te ga je potrebno izdvojiti iz ovog kalibracijskog skupa uzoraka PMPS. Nakon uklanjanja ovoga netipičnog uzorka ponovljena je PLS-DA analiza na kalibracijskom setu uzoraka PMPS A i C, ali bez ovog netipičnog uzorka.



Slika 230. Raspodjela faktorskih bodova PC 1 i PC 2 matematički obrađenih (SNV i uklanjanje trenda polinomom 4. stupnja) Raman spektara PMPS A i C u području $\tilde{\nu}$ = 3500 - 100 cm^{-1} nakon uklanjanja netipičnog uzorka.

Kako bi odredili najznačajnije valne brojeve, načinjen je dijagram opterećenja (Slika 231.).

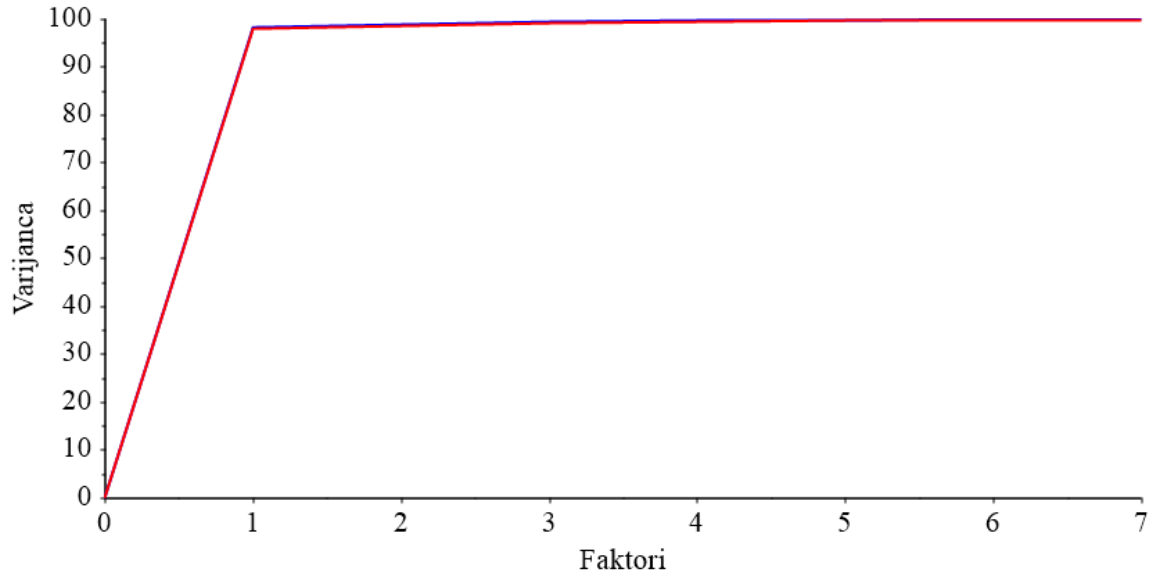


Slika 231. Profil pterećenja za jednu latentnu varijablu (LV) po valnim brojevima ($\tilde{\nu}$).

Na Slici 231. jasno su vidljive vrpce odgovorne za razdvajanje dviju klasa unutar modela, odnosno razdvajanje meningokoknih polisaharida serogrupe A i C. To su karakteristične spektralne vrpce pri: $\tilde{\nu}$ = 2980 cm^{-1} koje proizlaze od C-H i CH_2 istezanja; $\tilde{\nu}$ = 2930 cm^{-1} proizlaze od CH_2 asimetričnog istezanja; $\tilde{\nu}$ = 2875 cm^{-1} proizlaze od CH_2 simetričnog istezanja; $\tilde{\nu}$ = 1750 - 1700 cm^{-1} C=O vibracija istezanja karbonilne grupe te od $\tilde{\nu}$ = 1680 - 1639 cm^{-1} proizlaze od C=ONHR istezanja; $\tilde{\nu}$ = 1500 - 1200 cm^{-1} C-H/CH_2 deformacijske vibracije; $\tilde{\nu}$ = 1350 - 1140 cm^{-1} P=O istezanja; $\tilde{\nu}$ = 1150 - 1070 cm^{-1} proizlazi od C-O-C asimetričng istezanja; $\tilde{\nu}$ = 980 - 965 cm^{-1} proizlaze od C-H savijanja van ravnine; $\tilde{\nu}$ = 800 - 100 cm^{-1} deformacijske vibracije C-C-O grupa (Larkin, 2018).

4.3.8.1.Optimizacija Raman PLS-DA modela

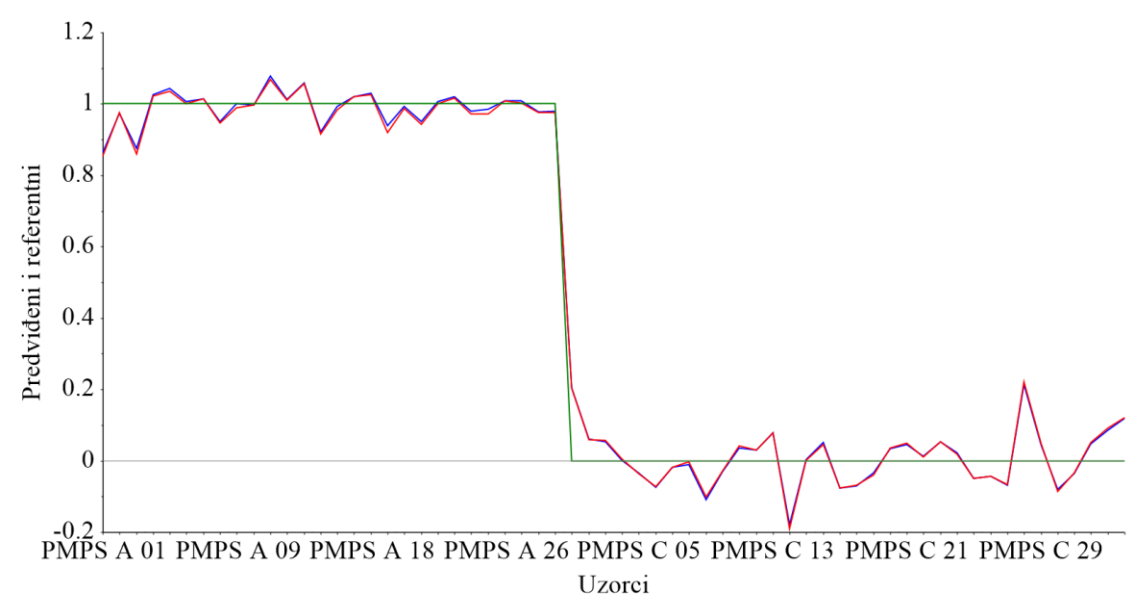
Kako bi odredili optimalni broj PLS faktora, bilo je potrebno prikazati kumulativnu kalibracijsku i validacijsku varijancu za svaki PLS faktor (Slika 232.).



Slika 232. Kumulativna kalibracijska i validacijska varijanca za svaki PLS faktor. Kalibracija (plava linija), validacija (crvena linija).

Grafikon kumulativne varijance pokazuje 98,2 % kalibracijske i 98,1 % validacijske varijance za jedan faktor. Ovaj mali broj PLS faktora sugerira nisku korelaciju u spektrima različitih klasa, ali sličnosti u spektrima unutar klase.

Optimizacija PLS-DA modela provedena je pomoću trening seta uzoraka postupkom unakrsne validacije.

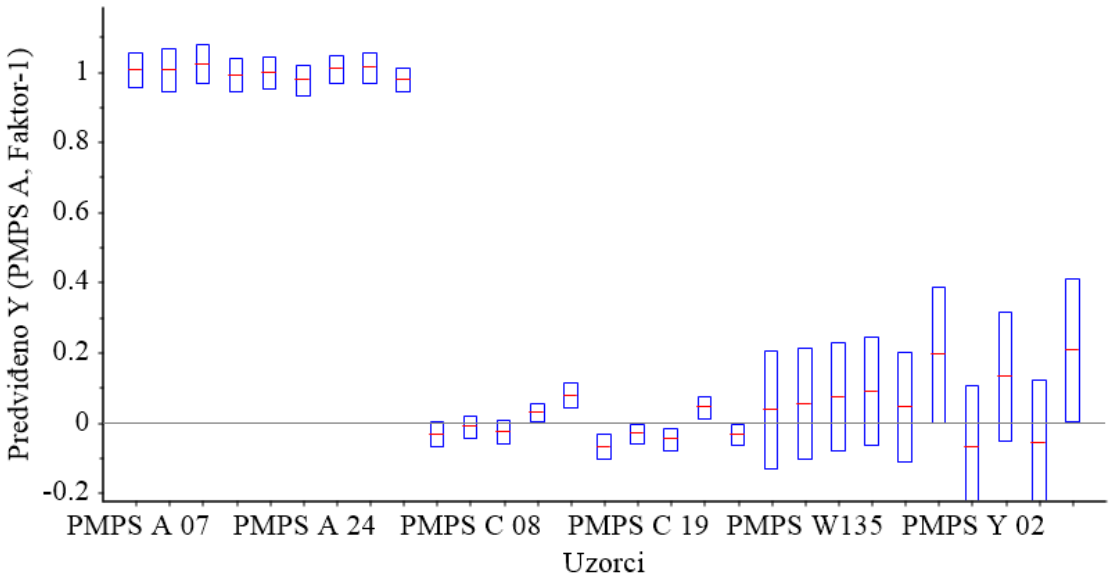


Slika 233. Odnos predviđenih i referentnih vrijednosti uzoraka PMPS A i PMPS C. Kalibracija (plava linija), validacija (crvena linija), refrentna vrijednost (zelena linija).

Slika 233. prikazuje jasno odvajanje među uzorcima PMPS A i PMPS C. Validacijski uzorci preklapaju se sa kalibracijskim uzorcima ispravno i u skladu s odgovarajućom klasom (serogrupom) meningokoknih polisaharida.

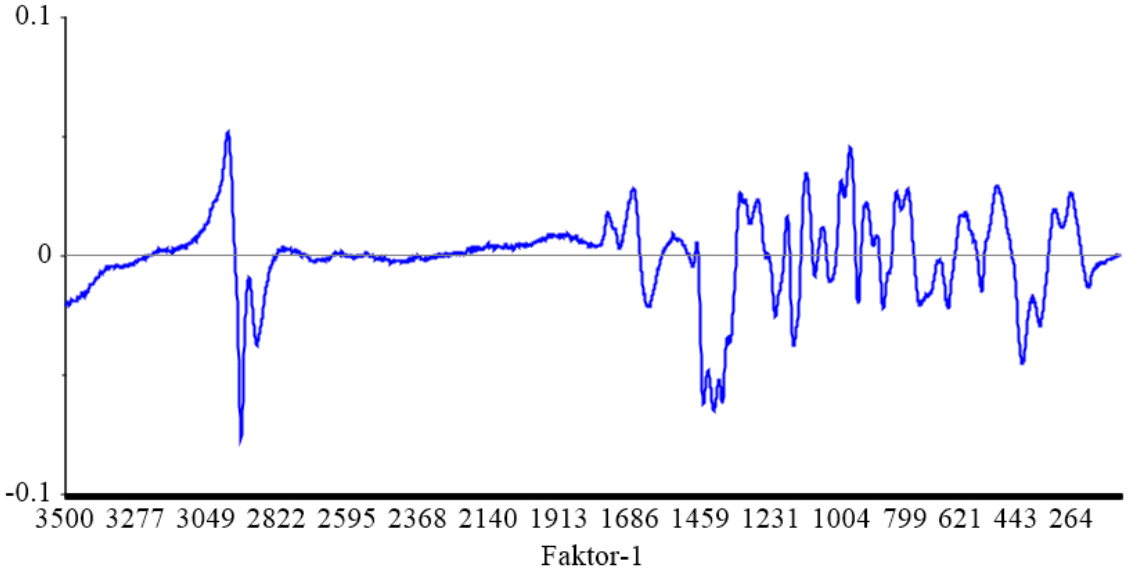
4.3.8.2.Validacija Raman PLS-DA modela

Kako bi validirali PLS-DA model s jednim PLS faktorom, provedena je identifikacija nepoznatih uzoraka PMPS A i C iz vanjskog test seta. Rezultati identifikacije ovim modelom prikazani su na Slici 234.



Slika 234. Predviđene vrijednosti uzoraka vanjskog test seta s procjenjenim odstupanjem dobivene formiranim PLS-DA modelom sa jednim PLS faktorom.

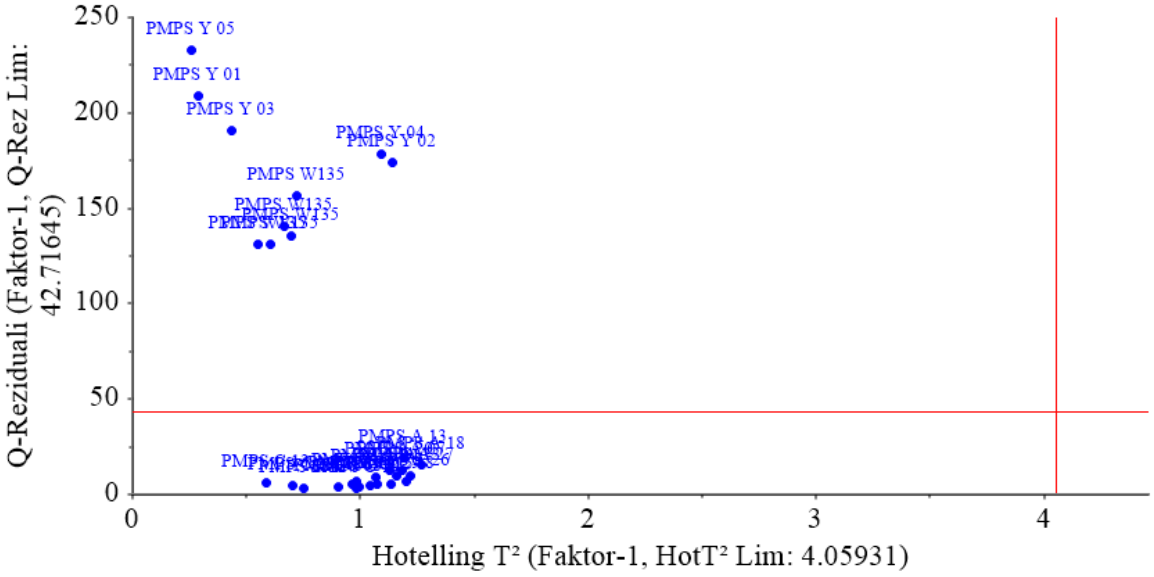
Slika 234. prikazuje kako PLS-DA model sa jednim PLS faktorom identificira nepoznate uzorke PMPS A i C iz vanjskog test seta. Ovdje je jasno da su uzorci PMPS A raspodijeljeni oko idealne vrijednosti 1 i ovi su uzorci sasvim ispravno dodjeljeni svojoj klasi. Očekivano su svi uzorci PMPS W135 i Y pridruženi klasi C. Svi uzorci PMPS C ispravno su dodjeljeni svojoj klasi. Iz ovdje opisanih rezultata može se zaključiti da PLS-DA model sa jednim PLS faktorom ima izuzetno dobru sposobnost klasifikacije PMPS A i C. Uzorci PMPS W135 i PMPS Y, koji su i ovdje korišteni kao negativna proba, dodjeljeni su klasi PMPS C što je bilo i za očekivati obzirom na njihovu sličnost u kemijskoj strukturi.



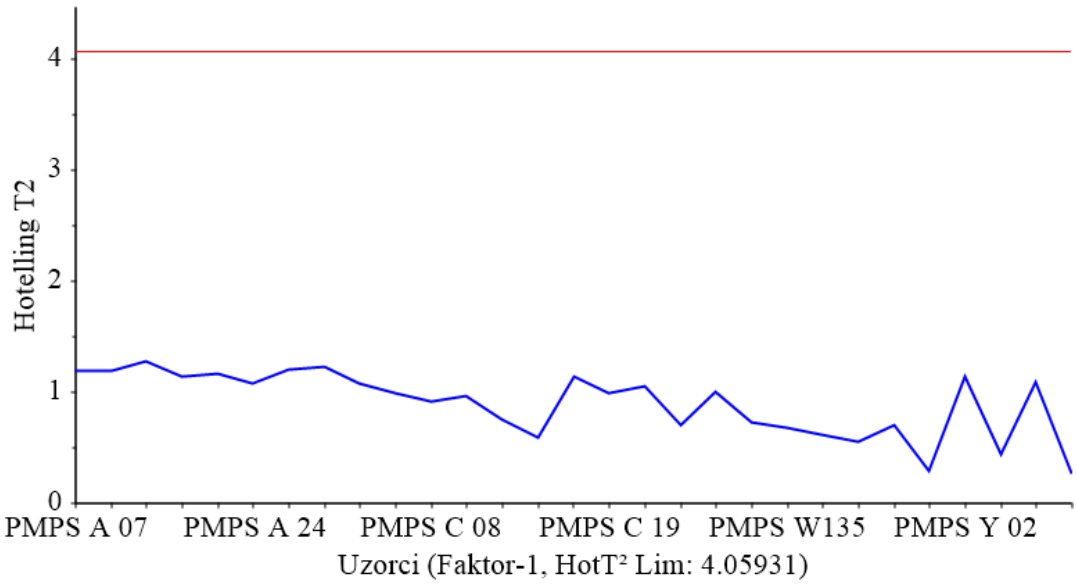
Slika 235. Profil pterećenja za jednu latentnu varijablu (LV) po valnim brojevima ($\tilde{\nu}$).

Na Slici 235. jasno su vidljive vrpce odgovorne za razdvajanje dviju klasa unutar modela, odnosno razdvajanje meningokoknih polisaharida serogrupe A i C. To su karakteristične spektralne vrpce pri: $\tilde{\nu}$ = 2980 cm^{-1} koje proizlaze od C-H i CH_2 istezanja; $\tilde{\nu}$ = 2930 cm^{-1} proizlaze od CH_2 asimetričnog istezanja; $\tilde{\nu}$ = 2875 cm^{-1} proizlaze od CH_2 simetričnog istezanja; $\tilde{\nu}$ = 1750 - 1700 cm^{-1} C=O vibracija istezanja karbonilne grupe te od $\tilde{\nu}$ = 1680 - 1639 cm^{-1} proizlaze od C=ONHR istezanja; $\tilde{\nu}$ = 1500 - 1200 cm^{-1} C-H/CH_2 deformacijske vibracije; $\tilde{\nu}$ = 1350 - 1140 cm^{-1} P=O istezanja; $\tilde{\nu}$ = 1150 - 1070 cm^{-1} proizlazi od C-O-C asimetričng istezanja; $\tilde{\nu}$ = 980 - 965 cm^{-1} proizlaze od C-H savijanja van ravnine; $\tilde{\nu}$ = 800 - 100 cm^{-1} deformacijske vibracije C-C-O grupa (Larkin, 2018).

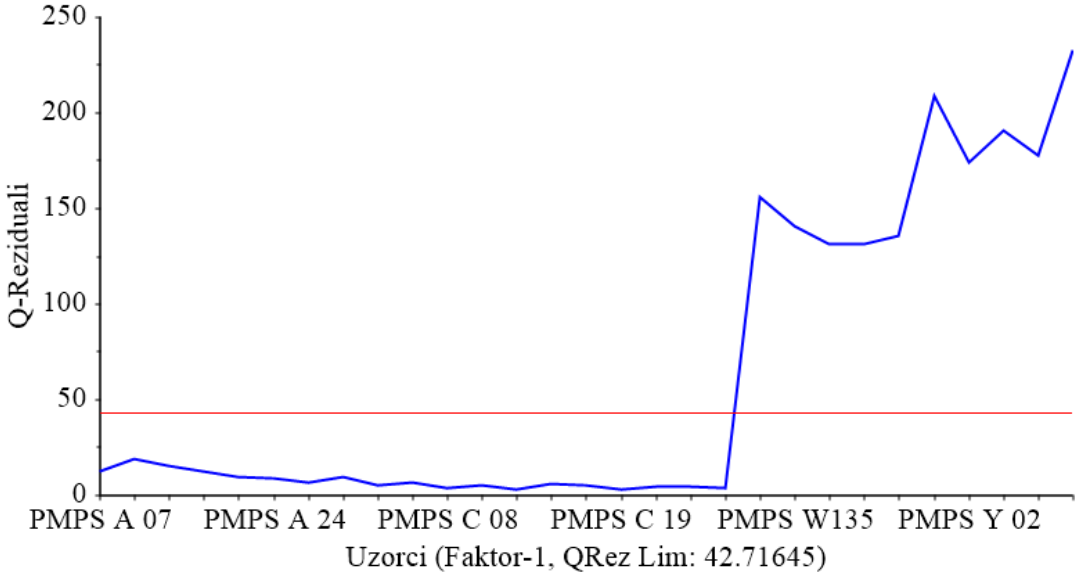
Prisutnost netipičnih i ekstremnih uzoraka u modelu može se procjeniti pomoću Hotelling T^2 statistike te Q reziduala. Kako PLS-DA model pruža aproksimativno rješenje, Q statistika se koristi za procjenu usklađenosti svakog uzorka s modelom. Hotelling T^2 statistika te Q reziduali proveli su se u svrhu potvrde identifikacije uzoraka PMPS W135 i PMPS Y kao netipičnih uzoraka.



Slika 236. Hotelling T^2 statistika i Q-Reziduali uzoraka sa pripadajućom kritičnom vrijednosti (crvena linija).



Slika 237. Hotelling T^2 statistika sa pripadajućom kritičnom vrijednosti (crvena linija).



Slika 238. Q reziduali uzoraka s pripadajućom graničnom vrijednosti (crvena linija).

Na Slikama 236. - 238. su vidljivi uzorci meningokoknih polisaharida serogrupa W135 i Y koji su PLS-DA modelom prepoznati kao potencijalni netipični uzorci. Dobiveni su očekivani rezultati gdje su uzorci negativne probe dodjeljeni klasi meningokoknih polisaharida serogrupe C radi sličnost u kemijskoj strukturi. Kod dentifikacije PLS-DA modelom ispitivani uzorci uvijek se dodjeljuju jednoj od klasa. Budući su u Imunološkom zavodu serogrupe meningokoknih polisaharida W135 i Y samo eksperimentalno proizvedene i ne proizvode se redovnom proizvodnjom niti sudjeluju u redovnoj kontroli kvalitete, nema nikakvih mogućnosti da se tijekom redovite analize uzoraka PMPS pogrešno identificiraju kao PMPS C.

Za procjenu prediktivne sposobnosti Raman PLS-DA modela korišteni su ovi parametri: osjetljivost, specifičnost i učinkovitost.

Tablica 15. Matrica zabune validacijskih parametara za PMPS uzorke iz vanjskog validacijskog seta dobivenih Raman PLS-DA modelom s jednim PLS faktorom.

stvarno/predviđeno	klasa PMPS A	klasa PMPS C	nije klasificirano
klasa PMPS A	9	0	0
klasa PMPS C	0	10	0
klasa PMPS W135	0	5	0
klasa PMPS Y	0	5	0
CSNS	100%	100%	TSNS = 100%
CSPS	100%	0%	TSPS = 0%
CEFF	100%	0%	TEFF (jedan PLS) = 0%

CSNS, CSPS, CEFF, TSNS, TSPS i TEFF su opisani u poglavlju 2.5.3.2.

Ukoliko se uzmu u obzir ovdje dobiveni rezultati, može se zaključiti da je razvijeni Raman PLS-DA model visoko učinkovit za identifikaciju novih uzoraka PMPS A i C.

5. **ZAKLJUČCI**

Na temelju rezultata prikazanih u ovome doktorskome radu može se zaključiti, kako slijedi:

1. U ovome je radu detaljno istraženo i precizno utvrđeno da se primjenom novih, brzih i nedestruktivnih metoda vibracijske spektroskopije - spektroskopijom bliskoga infracrvenoga zračenja (NIR) i Ramanovom spektroskopijom, može kroz kratak vremenski period prikupiti vrlo veliki skup podataka i, u kombinaciji s kemometrijskim metodama, uspješno diskriminirati različiti pročišćeni meningokokni polisaharidi određenih serogrupa (PMPS). Iz snimljenih NIR i Ramanovih spektara proizvodnih serija PMPS A i C kao i eksperimentalnih serija PMPS W135 i Y (koji su bili negativne probe, a čija je kemijska struktura slična kemijskoj strukturi PMPS C), eksploracijskom analizom glavnih komponenti (PCA) grupirani su slični i jasno razdvojeni različiti spektralni podaci.

2. Primjenom dvaju kemometrijskih alata za klasifikaciju - (1) mekog neovisnog modeliranja analogne klase (SIMCA) i (2) diskriminantne analize parcijalnih najmanjih kvadrata (PLS-DA) formirani su NIR i Raman SIMCA i PLS-DA modeli. Svi formirani modeli (NIR SIMCA i NIR PLS-DA, Raman SIMCA i Raman PLS-DA) su uspješno optimizirani i validirani uz korištenje spektara PMPS A i C kao i negativnih proba (PMPS W135 i Y) te je utvrđena valjanost formiranih i validiranih NIR i Raman modela, koji uspješno klasificiraju 100 % nepoznatih uzoraka PMPS A i C. Dodatno, oba NIR modela, koji su formirani na temelju podataka o proizvodnim serijama PMPS A i C, uspješno su potvrdili identitet NIBSC standarada ovih dvaju polisaharida. Utvrđene su i najutjecajnije spektralne regije PMPS A i C na pojedinu glavnu komponentu.

3. NIR i Raman pristupom SIMCA jednoklasnog modeliranja također su uspješno istraženi zasebni pouzdani autentifikacijski modeli za PMPS A i PMPS C. Primjena ovoga tipa klasifikacijskih modela se preporuča u proizvodnji cjepiva ali i općenito u biotehnološkoj proizvodnji i to u rješavanju iznimno važnih autentifikacijskih problema uzoraka kod kojih su, osim ciljnih, obično prisutni i drugi neciljni uzorci.

4. Uspješna primjena NIR i Raman modela za identifikaciju PMPS A i C je u skladu s Direktivom 2010/63/EU i izvrsna je i ekonomski učinkovita alternativa referentnoj dvostrukoj imunodifuziji - Ouchterlony metodi u uporabi.

5. Kvalitativni istraživački pristup i rezultati opisani u ovom doktorskom radu mogu se primijeniti u razvoju novih NIR i Raman modela za identifikaciju drugih polisaharida, kao što su meningokokni polisaharidi serogrupa B, W135 i Y. Dodatno, ovdje opisani rezultati jasno upućuju na potencijal metoda vibracijske spektroskopije u kombinaciji sa multivarijantnim tehnikama za analizu složenih bioloških matriksa i klasifikacijsku primjenu, kako u farmaceutskoj tako i u biotehnološkoj industriji, osobitu u autentifikacijske svrhe.

6. LITERATURA

1. Abdi, H. and Williams, L. J. (2010) Principal component analysis. *WIREs Comp Stats*, **2**, 433–459.

2. Abonyi, J., Feil, B. (2007) *Cluster Analysis for Data Mining and System Identification, Cluster Analysis for Data Mining and System Identification*. Birkhäuser Basel, Basel.

3. Adams, M.J. (2004) *Chemometrics in Analytical Spectroscopy*, 2. izd., Royal Society of Chemistry, Cambridge.

4. Afseth, N.K., Kohler, A. (2012) Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. *Chemom. Intell. Lab. Syst.* **117**, 92–99.

5. Arcos, M.J., Ortiz, M.C., Villahoz, B., Sarabia, L.A. (1997) Genetic-algorithm-based wavelength selection in multicomponent spectrometric determinations by PLS: Application on indomethacin and acemethacin mixture. *Anal. Chim. Acta* **339**, 63–77.

6. Balabio, Davide and Todeschini, R. (2009*) Infrared Spectroscopy for Food Quality Analysis and Control: Multivariate Classification for Qualitative Analysis*, Elsevier/Academic Press, Amsterdam /London, str. 83–104.

7. Ballabio, D., Consonni, V. (2013) Classification tools in chemistry. Part 1: linear models. PLS-DA. *Anal. Methods* **5**, 3790-3798.

8. Ballabio, D., Grisoni, F., Todeschini, R. (2018) Multivariate comparison of classification performance measures. *Chemom. Intell. Lab. Syst.* **174**, 33–44.

9. Barker, M., Rayens, W. (2003) Partial least squares for discrimination. *J. Chemom*. **17**, 166–173.

10. Barnes, R.J., Dhanoa, M.S., Lister, S.J. (1989) Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra. *Appl. Spectrosc*. **43**, 772–777.

11. Bartholomew, D.J. (2010) *Analysis and interpretation of multivariate data.* Int. Encycl. Educ. Elsevier/London, str 12–17.

12. Berrueta, L.A., Alonso-Salces, R.M., Héberger, K. (2007) Supervised pattern recognition in food analysis. *J. Chromatogr. A* **1158**, 196–214.

13. Biancolillo, A., Marini, F. (2018) Chemometric methods for spectroscopy-Based pharmaceutical analysis. *Front. Chem.* **6**, 576.

14. Blanco, M., Villarroya, I. (2002) NIR spectroscopy: A rapid-response analytical tool. *TrAC - Trends Anal. Chem.* **21**, 240–250.

15. Bocklitz, T., Walter, A., Hartmann, K., Rösch, P., Popp, J. (2011) How to pre-process Raman spectra for reliable and stable models? *Anal. Chim. Acta* **704**, 47–56.

16. Bratchell, N. (1992), *Multivariate Pattern Recognition in Chemometrics*, Elsevier Science, Amsterdam.

17. Brereton, R.G. (2018) *Chemometrics: Data Driven Extraction for Science*: *Principal Component Analysis and Unsupervised Pattern Recognition*, 2. izd., John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, 163–214.

18. Brereton, R.G. (2015) Pattern recognition in chemometrics. *Chemom. Intell. Lab. Syst.* **149**, 90–96.

19. Brereton, R.G., Jansen,, J., Lopes, Marini, F., Pomerantsev, A., Rodionova, O., Roger, J.M., Walczak,, B., Tauler, R. (2018) Chemometrics in analytical chemistry—part II: modeling, validation, and applications, *Anal. Bioanal. Chem.* **410** (26): 6691–6704.

20. Brereton, R.G., Jansen, J., Lopes, J., Marini, F., Pomerantsev, A., Rodionova, O., Roger, J.M., Walczak, B., Tauler, R.. (2017) Chemometrics in analytical chemistry—part I: history, experimental design and data analysis tools. *Anal. Bioanal. Chem.* **409**, 5891–5899.

21. Brereton, R.G., Lloyd, G.R. (2014) Partial least squares discriminant analysis: Taking the magic away. *J. Chemom.* **28**, 213–225.

22. Bro, R., Smilde, A.K. (2014) Principal component analysis. *Anal. Methods* **6**, 2812–2831.

23. Burns, D.A., Ciurczak, E.W. (2008) *Handbook of Near-Infrared Analysis*, 3. izd., Marcel Dekker, Inc, New York.

24. Burns, D.A., Ciurczak, E.W. (2001) *Handbook of Near-Infrared Analysis*, 2. izd., Marcel Dekker, Inc, New York.

25. Camo Analytics (2014) The Unscrambler X v 10.3, Camo Analytics AS, Oslo, Norveška <http://www.camo.com/downloads/U9.6%20pdf%20manual/The%20Unscrambler%20Methods.pdf_ >. Pristupljeno 21. travnja 2020. godine.

26. Cid, M.-M., Bravo, J. (2015) *Structure Elucidation in Organic Chemistry: The Search for the Right Tools*, Wiley-VCH Verlag GmbH & Co. KgaA.

27. CFR (2019) Code of Federal regulations (2019), Title 21, Volume 7, 1¸ Sec. 610.14 – Identity.

28. De Beer, T., Burggraeve, A., Fonteyne, M., Saerens, L., Remon, J.P., Vervaet, C.(2011) Near infrared and Raman spectroscopy for the in-process monitoring of pharmaceutical production processes. *Int J Pharm.* **417**, 32-47

29. De Luca, S., Bucci, R., Magrì, A.D., Marini, F. (2018) *Encyclopedia of Analytical Chemistry*: *Class Modeling Techniques in Chemometrics: Theory and Applications*. John Wiley & Sons, Ltd. New Jersey, str.1–24.

30. Dubes, R.C., Jain A K. (1988) *Algorithms for Clustering Data*, Prentice Hall.Inc. New Jersey

31. Einax, J.W., Truckenbrodt, D., Kampe, O. (1998) River Pollution Data Interpreted by Means of Chemometric Methods. *Microchem. J.* **58**, 315–324.

32. Ermer, J. (2015). *Method performance characteristics. Method Validation in Pharmaceutical Analysis. A Guide to Best Practice*. 2nd Edition, Wiley VCH, Weinheim, Germany, 73-182.

33. Esbensen, K.H., Geladi, P. (2010) Principles of proper validation: Use and abuse of re-sampling for validation. *J. Chemom.* **24**, 168–187.

34. Esbensen, K.H., Geladi, P. (2009) Principal Component Analysis: Concept, Geometrical Interpretation, Mathematical Background, Algorithms, History, Practice. *Compr. Chemom.* **2**, 211–226.

35. EP (2010) Directive 2010/63/EU - On the protection of animals used for scientific purposes. Off. J. Eur. Union. 33–79. EP, European Parliament, Bruxelles, https://doi.org/32010L0063. Pristupljeno 21.ožujka 2021.

.

36. Ph.Eur. (2019) Meningococcal polysaccharide vaccine, Ph. Eur. - European Pharmacopoeia, European Council of Europe, Strasbourg.

37. Ferreira, L., Hitchcock, D.B. (2009) A comparison of hierarchical methods for clustering functional data. *Commun. Stat. Simul. Comput*. **38**, 1925–1949.

38. Gabutti, G., Stefanati, A., Kuhdari, P. (2015) Epidemiology of Neisseria meningitidis infections: Case distribution by age and relevance of carriage. *J. Prev. Med. Hyg*. **56(3),** E116–E120.

39. Geladi, Paul and Kowalski, B.R. (1986) Partial least squares regression: A tutorial. *Anal. Chim. Acta* **185**, 1–17.

40. Geladi, P. (2003) Chemometrics in spectroscopy. Part 1. Classical chemometrics. *Spectrochim. Acta - Part B At. Spectrosc*. **58**, 767–782.

41. Geladi, P., MacDougall, D., Martens, H. (1985) Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat. *Appl. Spectrosc*. **39**, 491–500.

42. Görög, S. (2015) Identification in drug quality control and drug research. *TrAC - Trends Anal. Chem*. **69**, 114–122.

43. Gotschlich, E.C., Goldschneider, I., Artenstein, M.S. (1969a) Human immunity to the

meningococcus. IV. Immunogenicity of group A and group C meningococcal polysaccharides in human volunteers. *J. Exp. Med*. **129,** 1367–84.

44. Gotschlich, E.C., Liu, T.Y., Artenstein, M.S. (1969b) Human immunity to the meningococcus:.III. Preparation and immunochemical properties of the group A, group B, and group C meningococcal polysaccharides. *J. Exp. Med*. **129 (6**), 1349–1365.

45. Harschel, W. (1800) Investigation of the powers of the prismatic colours to heat and illuminate objects; with remarks, that prove the different refrangibility of radiant heat. To which is added, an inquiry into the method of viewing the sun advantageously, with telescopes of. *Philos. Trans. Roy. Soc. Lon* **90**, 255–283.

46. Hopke, P.K. (2003) The evolution of chemometrics. *Anal. Chim. Acta* **500**, 365–377.

47. Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24**, 417–441.

48. Huang, J, Romero-Torres, S, Moshgbar, M (2010) Practical considerations in data pre-treatment for NIR and Raman spectroscopy, *American Pharmace. Rev*, **13(6),**116-127.

49. Jackson, J. E, Mudholkar, G. S. (1979) Control Procedures for Residuals Associated With Principal Component Analysis, *Technometrics,* **21**, 341-349.

50. Jolliffe, T.I. (2002) *Principal components analysis*, Springer-Verlag New York.

51. Kabat,.E.A., Kaiser, H., Sikorski, H. (1944) Preparation of the type-specific polysaccharide of the type I meningococcus and a study of its efectiveness as an antigen in human beings. *J. Exp. Med* **80(4),** 299–307.

52. Kabat, E.A., Bezer, A.E. (1958) The effect of variation in molecular weight on the antigenicity of dextran in man. *Arch. Biochem. Biophys*. **78**, 306–18.
53. Kennard, R.W., Stone, L.A. (1969) Computer Aided Design of Experiments. *Technometric* **11**, 137–148.

54. Kohler, A., Afseth, N., Martens, H. (2010) *Handbook of Vibrational Spectroscopy: Chemometrics in Biospectroscopy*, John Wiley & Sons, Ltd., New Jersey.

55. Kumar, N., Bansal, A., Sarma, G.S., Rawal, R.K. (2014) Chemometrics tools used in analytical chemistry: An overview. *Talanta* **123**, 186–199.

56. Landberg,G., Mendelstam, L. (1928) Eine neue Erscheinung bei der Lichtzerstreuung in Krystallen. *Naturwissenschaften* **16**, 557–558.

57. Larkin, P. J. (2018) *Infrared and Raman Spectroscopy*, 2. izd., Elsevier Inc.Amsterdam.

58. Lopez, M., I., Callao, M., P., Ruisanchez, I. (2015) A tutorial on the validation of qualitative methods: From the univariate to the multivariate approach. *Anal. Chim. Acta*. **891**, 62–72.

59. Luypaert, J., Massart, D.L., Vander Heyden, Y. (2007) Near-infrared spectroscopy applications in pharmaceutical analysis. *Talanta* **72**, 865–883.

60. Marini, F. (2013) *Chemometrics in Food Chemistry*, Elsevier, Amsterdam.

61. Martens, H., Jensen, S.A., G. (1983) *Proceedings of the Nordic Symposium on Applied Statistics: Multivariate linearity trans-formations for near infrared reflectance spectroscopy*, O.H.J.Christie, Stokkland Forlag, Stavanger, Norway, 205–234.

62. Martens, H., Nielsen, J.P., Engelsen, S.B. (2003) Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures. *Anal. Chem*. **75**, 394–404.

63. Massart, D.L., Vandeginste, B.G. (1998) *Handbook of Chemometrics and Qualimetrics*. Elsevier Science Inc ,New York.

64. Mitsutake, H., Poppi, R.J., Breitkreitz, M.C. (2019) Raman imaging spectroscopy: History, fundamentals and current scenario of the technique. *J. Braz. Chem. Soc*. **30**, 2243–2258.

65. Mujica, L.E., Rodellar, J., Fernandez, A., Guemes, A. (2010) Q-statistic and T2-statistic

PCa-based measures for damage assessment in structures. *Struct. Health Monit.* **10**, 539-553.

66. Murtagh, F., Contreras, P. (2012) Algorithms for hierarchical clustering: An overview. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov*. **2**, 86–97.

67. Naes, T., Isaksson, T.F., Davis, T. (2002) *A User Friendly Guide to Multivariate Calibration and Classification*, NIR Publications. Chichester.

68. Prakash, M.M., Dagaonkar, A.(2011) *Recent res. sci. technol* **3**, 41–50.

69. Oliveri, P. (2017) Class-modelling in food analytical chemistry: Development, sampling, optimisation and validation issues – A tutorial. *Anal. Chim. Acta* **982**, 9–19.

70. Oliveri, P., Downey, G. (2012) Multivariate class modeling for the verification of food-authenticity claims. *TrAC - Trends Anal. Chem*. **35**, 74–86.

71. Panatto, D., Amicizia, D., Lai, P.L., Cristina, M.L., Domnich, A., Gasparini, R. (2013) New versus old meningococcal Group B vaccines: How the new ones may benefit infants & toddlers. *Indian J. Med. Res*. **138(6),** 835-846.

72. Pasquini, C. (2003) Near infrared spectroscopy: Fundamentals, practical aspects and analytical applications. *J. Braz. Chem. Soc*. **14**, 198–219.

73. Pearson K. (1901) On lines and planes of closest fit to systems of points in space. *Philos. Mag*. **2**, 559–572.

74. Pomerantsev, A.L.(2014) *Chemometrics in Excel*. John Wiley & Sons, Inc., Hoboken, New Jersey.

75. Pomerantsev, A.L., Rodionova, O.Y. (2020) Popular decision rules in SIMCA: Critical review. *J. Chemom*. **34**, 1–14.

76. Pomerantsev, A.L., Rodionova, O.Y. (2018) Multiclass partial least squares discriminant analysis: Taking the right way—A critical tutorial. *J. Chemom.* **32**, 1–16.

77. Pomerantsev, A.L., Rodionova, O.Y. (2014) Concept and role of extreme objects in PCA/SIMCA. *J. Chemom.* **28**, 429–438.

78. Pravdova, V., Walczak, B., Massart, D.L., Kawano, S., Toyoda, K., Tsenkova, R. (2001) Calibration of somatic cell count in milk based on near-infrared spectroscopy. *Anal. Chim. Acta* **450**, 131–141.

79. Qin, S.J. (2003) Statistical process monitoring: Basics and beyond. *J. Chemom.* **17**, 480–502.

80. Raman, C. V., Krishnan, K.S. (1928) A new type of secondary radiation. *Nature* **121**, 501–502.

81. Randriamihamison, N., Vialaneix, N., Neuvial, P. (2020) Applicability and Interpretability of Ward's Hierarchical Agglomerative Clustering With or Without Contiguity Constraints. *J. Classif.* (objavljeno *online* 30.rujna.2020). doi: https://doi.org/10.1007/s00357-020-09377-y.

82. Riedl, J., Esslinger, S., Fauhl-Hassek, C. (2015) Review of validation and reporting of non-targeted fingerprinting approaches for food authentication. *Anal. Chim. Acta*. **885**, 17-32.

83. Rinnan, Å., Berg, F. van den, Engelsen, S.B. (2009) Review of the most common pre-processing techniques for near-infrared spectra. *TrAC - Trends Anal. Chem.* **28**, 1201–1222.

84. Rodionova, O.Y., Balyklova, K.S., Titova, A. V., Pomerantsev, A.L. (2014) Quantitative risk assessment in classification of drugs with identical API content. *J. Pharm. Biomed. Anal.* **98**, 186–192.

85. Rodionova, O.Y., Pomerantsev, A.L. (2020) Chemometric tools for food fraud detection: The role of target class in non-targeted analysis. *Food Chem*. **317**, 126448.

86. Rodionova, O.Y., Pomerantsev, A.L. (2006) Chemometrics: achievements and prospects. *Russ. Chem. Rev.* **75**, 271–287.

87. Rodionova, O.Y., Titova, A. V., Balyklova, K.S., Pomerantsev, A.L. (2019) Detection of counterfeit and substandard tablets using non-invasive NIR and chemometrics - A conceptual framework for a big screening system. *Talanta* **205**, 120150.

88. Rodionova, O.Y., Titova, A. V., Pomerantsev, A.L. (2016) Discriminant analysis is an inappropriate method of authentication. *TrAC - Trends Anal. Chem*. **78**, 17–22.

89. Rogers, D., Hopfinger, A.J. (1993) Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. *J. Chem. Inf. Comput. Sci*. **33**, 854–866

.

90. Šašić, S. (2007) *Pharmaceutical Applications of Raman Spectroscopy,* John Wiley & Sons, Inc., Hoboken, New Jersey.

91. Scherp H.W., Rake G. (1945) Studies on meningococcal infection: XIII. Correlation between antipolysaccharide and the antibody which protects mice against infection with type I meningococci, *J.Exp.Med* **81(1),** 85-92.

92. Slišković, D., Grbić, R., Hocenski, Ž., (2012) Multivariate statistical process monitoring.*Tehnički vjesnik* **19**, 33-41.

93. Smekal A., (1928) Zur Quantentheorie der Streuung und Dispersion. *Naturwissenschaften* **16**, 612–613.

94. The United State Pharmacopoeia, USP 40-NF 35 (2017), 1090 Assessment ofdrug product performance - bioavailability, bioequivalence and dissolution, U.S.P Pharmacopoeial Convention, Rockville.

95. Ward, J.H. (1963) Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **58**, 236-244.

96. Weichselbaum, A. (1887) Ueber die Aetiologie der akuten meningitis cerebrospinalis. *Fortschr Med*. **5**, 573–583.

97. WHO (1999) Group A and C meningococcal vaccines : WHO position paper. WHO, World Health Organisation, Geneve https://apps.who.int/iris/bitstream/handle/10665/230909/WER7436_297-303.PDF?sequence=1&isAllowed=y. Pristupljeno, 21. ožujka 2021.

98. WHO (2014) Technical Report Series (TRS), No. 594 (1975) Requirements for Meningococcal Polysaccharide vaccine, Annex 2, WHO, World Health Organisation, Geneve.

99. Wiberg, K. (2004) Multivariate spectroscopic methods for the analysis of solutions. PhD Thesis. Department of Analytical Chemistry, Stockholm University, Stockholm, Sweden.

100. Willett, D.R., Rodriguez, J.D. (2018) Quantitative Raman assays for on-site analysis of stockpiled drugs *Anal. Chim. Acta* **1044**, 131–137.

101. Wise, B.M., Gallagher, N.B., Watts Butler, S., White, D.D., Barna, G.G. (1997) Development and Benchmarking of Multivariate Statistical Process Control Tools for a Semiconductor ETCH Process: Impact of Measurement Selection and Data Treatment on Sensitivity. *IFAC Proc*. **30**, 35–42.

102. Wold, S., Ruhe, A., Wold, H., Dunn, W.J. (2014) Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Siam. J.Sci.Stat. Comput.* **31**, 274–295.

103. Wold, S.,Ruhe,A.,Wold, H., Dunn,W.J. (1984) The Collinearity Problem in Linear Regression.The Partial Least Squares (PLS) Approach to Generalized Inverses. *Siam. J.Sci.Stat. Comput.* **5**, 735–743.

104. Wold, H. (1975) Soft Modelling by Latent Variables: The Non-Linear Iterative Partial Least Squares (NIPALS) *Approach. J. Appl. Probab*. **12**, 117–142.

105. Wold, S. (1995) Chemometrics; what do we mean with it, and what do we want from it? *Chemom. Intell. Lab. Syst.* **30**, 109–115.

106. Wold, S. (1976) Pattern recognition by means of disjoint principal components models.

Pattern Recognit. **8**, 127–139.

107. Wold, S., Esbensen, K., Geladi, P. (1987) Principal component analysis. *Chemom. Intellig. Lab. Syst*. **2**, 37–52.

108. Wold, S., Sjöström, M., Eriksson, L. (2001) PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst*. **58**, 109–130.

109. Zeaiter, M., Rutledge, D. (2009) Preprocessing Methods. *Compr. Chemom*. **3**, 121–231.

POPIS KRATICA I SIMBOLA

- CA - klasterska analiza - (engl. *Cluster Analysis*)
- *Confidence interval* - interval pouzdanosti
- *Cross validation* - unakrsna validacija
- EM zračenje - elektromagnetsko zračenje (engl. *Electromagnetic radiation*)
- ERV - objašnjena rezidualna varijanca (engl. *Explained Residual Variance*)
- ERVC - objašnjena rezidualna varijanca u kalibraciji (engl. *Explained Residual Variance in training*)
- ERVP - objašnjena rezidualna varijanca u validaciji engl. *Explained Residual Variance in validation*)
- FT - Fourierova transformacija (engl. *Fourier transformation*)
- IR spektroskopija - Infracrvena spektroskopija (engl. *Infrared spectroscopy*)
- *Leverages* - utjecajne vrijednosti
- *Loadings* - opterećenja
- *LV*- latentne varijable (engl. *latent variables*)
- MSC - korekcija višestrukog raspršenja (engl. *Multiplicative Scatter Correction*)
- NIR spektroskopija - spektroskopija bliskoga infracrvenog zračenja (engl. *Near infrared spectroscopy*)
- *Outlier* – netipični uzorak
- *Overfitting* - precijenjena sposobnost predviđanja
- PC - glavna komponenta (engl. *Principal Components*)
- PCA - analiza glavnih komponenata (engl. *Principal Component Analysis*)
- Ph.Eur - Europska Farmakopeja (engl. *European Pharmacopoeia*)
- PLS regresija - regresija parcijalnih najmanjih kvadrata (engl. *Partial Last Squares*)
- PLS-DA - diskriminantna analiza parcijalnih najmanjih kvadrata
- PMPS - pročišćeni meningokokni polisaharid (engl. *Partial least Squares Discriminant Analysis*)
- *Scores* - faktorski bodovi
- SIMCA - meko neovisno modeliranje analogne klase (engl. *Soft Independent Modelling of Class Analogy*)
- SNV - standardna normalna varijata (engl. *Standard Normal Variate*)
- *Underfitting* - podcijenjena sposobnost predviđanja
- USP-NF - Farmakopeja Sjedinjenih Država (USP) i Nacionalni formular (NF).

- WHO - Svjetska zdravstvena organizacija (eng. *World Health Organization*)
- WHO TRS - Serija tehničkih izvještaja svjetske zdravstvene organizacije (engl. *WHO Technical Report Series***)**
- λ – svojstvene vrijednosti (engl.*eigenvalues*)
- $\tilde{\nu}$ - valni broj

Životopis

Ana Mandac Zubak rođena je 1978. u Zadru. Osnovnu školu završila je u Zadru, a srednjoškolsko obrazovanje prirodoslovno-matematičkog usmjerenja u X. gimnaziji „Ivan Supek“, Zagreb. Školovanje je nastavila na Prehrambeno-biotehnološkom fakultetu Sveučilišta u Zagrebu na studiju Biotehnologija, smjer Biokemijsko inženjerstvo. Diplomski rad pod naslovom „*In vivo* istraživanja probiotičkog mehanizma djelovanja bakterije *Lactobacillus plantarum* L4“ izradila je u Laboratoriju za tehnologiju antibiotika, enzima, probiotika i starter kultura u Zavodu za biokemijsko inženjerstvo, pod mentorstvom prof. dr. sc. Jagode Šušković. Diplomirala je 2004. i stekla zvanje diplomiranog inženjera biotehnologije. Po završetku studija zaposlila se u Zavodu za javno zdravstvo Zadar u Odjelu za zdravstvenu ispravnost i kvalitetu voda, gdje je odradila svoj jednogodišnji pripravnički staž (rujan 2004. - rujan 2005.). Nakon uspješno završenog pripravničkog staža, zaposlila se na mjestu stručnog suradnika u Odsjeku za kemijsku kontrolu kvalitete, Imunološki zavod Zagreb (listopad 2005. - listopad 2014.). Poslijediplomski sveučilišni (doktorski) studij Biotehnologija i bioprocesno inženjerstvo upisala je 2010. na Prehrambeno-biotehnološkom fakultetu Sveučilišta u Zagrebu. Od listopada 2014. radi kao suradnik u Laboratoriju za polazne materijale, Hospira Zagreb d.o.o. Pfizer grupa. Kontinuirano se dodatno usavršava u zemlji i inozemstvu i to u području primjene analitičkih i statističkih metoda s naglaskom na validaciju i verifikaciju analitičkih metoda u okvirima koje definira Američka agencija za hranu i lijekove. Rezultati njenih istraživanja objavljeni su u dva rada iz skupine A1 u časopisima vrlo visokoga faktora odjeka (IF).

Znanstveni radovi iz skupine A1

Fabijanić, I., Čavužić, D., **Mandac Zubak, A.** (2019) Meningococcal polysaccharides identification by NIR spectroscopy and chemometrics. Carbohydrate polymers 216, 36-44. IF= 7,182 (Q1)

Mandac Zubak, A., Horvat, A., Čavužić, D., Fabijanić, I. (2020) Freeze-dried meningococcal vaccine: Total error assessment of a near-infrared method for water content determination. Talanta 211, 1-7. IF= 5,339 (Q1)