

**Többváltozós kalibrációs módszerek vizsgálata mint lehetséges
eszköz a Raman-térképezés alkalmazásához kontrollált
technológiákban**

OTDK-dolgozat

Készítette:

Nagy Brigitta

Témavezető:

Dr. Marosi György

Konzulens:

Farkas Attila

Budapesti Műszaki és Gazdaságtudományi Egyetem

Szerves Kémia és Technológia Tanszék

Budapest, 2014

A dolgozatban előforduló fontosabb rövidítések és betűszavak jegyzéke

ANN	Mesterséges ideghálózatok (<i>Artificial Neural Network</i>)
CCD	Töltéscsatolt eszköz, a Raman-spektroszkópiához használt detektor (<i>Charge-Coupled Device</i>)
CLS	Klasszikus legkisebb négyzetek módszere (<i>Classical Least Squares</i>)
GA	Genetikus algoritmusok, változókiválasztási módszer (<i>Genetic Algorithms</i>)
iPLS	Intervallum-PLS, változókiválasztási módszer
LWR	Helyileg súlyozott regresszió (<i>Locally Weighted Regression</i>)
MSC	Multiplikatív szóródási korrekció, adatelőkészítési módszer (<i>Multiplicative Scatter Correction</i>)
NIR	Közeli infravörös spektroszkópia (<i>Near Infrared Spectroscopy</i>)
PAT	Folyamatfelügyelő és Analizáló Technológia (<i>Process Analytical Technology</i>)
PCA	Főkomponens-analízis (<i>Principal Component Analysis</i>)
PCR	Főkomponens-regresszió (<i>Principal Component Regression</i>)
PLS	Részleges legkisebb négyzetek módszere (<i>Partial Least Squares</i>)
pPLS	Polinomiális PLS (<i>Polynomial PLS</i>)
RMSEC	A kalibráció közepes négyzetes hibája (<i>Root mean square error of calibration</i>)
RMSECV	A keresztellenőrzés közepes négyzetes hibája (<i>Root mean square error of cross validation</i>)
RMSEP	Az előrejelzés közepes négyzetes hibája (<i>Root mean square error of prediction</i>)
SNV	Sorstandardizálás, adatelőkészítési módszer (<i>Standard normal variate</i>)
SRD	Rangszámkülönbségek abszolút-érték összegének módszere (<i>Sum of Ranking Differences</i>)
SVM/ SVR	Támogatóvektorok-módszere/ Támogatóvektor-módszerrel végzett regresszió (<i>Support Vector Machine/ Support Vector Regression</i>)

Tartalomjegyzék

1. Bevezetés.....	5
2. Szakirodalmi áttekintés.....	7
2.1. A Raman-spektrometria kialakulása	7
2.2. Kémiai térképezés és alkalmazásai a gyógyszeriparban	8
2.2.1. Polimorf keverékek mennyiségi meghatározása	11
2.2.2. Hatóanyag-segédanyag rendszerek kvantitatív meghatározása.....	13
2.3. Ritkábban alkalmazott kemometriai módszerek a mennyiségi vizsgálatokban	15
3. Felhasznált anyagok és módszerek.....	18
3.1. A kísérletek során felhasznált anyagok és mintaelőkészítési módszerek	18
3.1.1. Polimorf rendszer mennyiségi elemzése	18
3.1.2. Hatóanyagot és segédanyagokat tartalmazó keverékek mennyiségi elemzése	19
3.2. A kísérletek során használt berendezések és vizsgálati körülmények.....	20
3.2.1. A berendezések adatai	20
3.2.2. A minták vizsgálati körülményei.....	23
3.2.2.1. <i>Polimorf rendszer mennyiségi elemzése</i>	23
3.2.2.2. <i>Hatóanyagot és segédanyagokat tartalmazó keverékek mennyiségi elemzése</i>	24
3.3. A számításokhoz alkalmazott szoftverek	24
3.4. A modellek validálására használt módszerek és a modelljóságot jellemző paraméterek....	25
3.5. Alkalmazott adatelőkészítési módszerek	27
3.6. Alkalmazott regressziós módszerek.....	29
3.7. Homogenitás-vizsgálat a makropixel-elemzés módszerével [49,50]	35
4. Eredmények	36
4.1. Polimorf rendszer mennyiségi elemzése.....	36
4.1.1. Adatelőkészítési módszerek vizsgálata	36
4.1.2. Referenciamódszerekkel felállított modellek	37
4.1.3. Változókiválasztási módszerekkel támogatott mennyiségi vizsgálatok	42
4.1.4. Nemlineáris regressziós módszerek tesztelése	44
4.1.5. A modellek összehasonlítása	48
4.1.6. A térképméret homogenításra gyakorolt hatásának vizsgálata	52
4.2. Hatóanyagot és segédanyagokat tartalmazó keverékek mennyiségi elemzése	54

4.2.1.	Az optimális térképméret meghatározása.....	54
4.2.2.	Adatelőkészítési módszerek vizsgálata	55
4.2.3.	A modellek validálása	57
4.2.4.	Referenciamódszerekkel elért eredmények.....	60
4.2.5.	Változókiválasztással támogatott PLS regresszió	62
4.2.6.	Nemlineáris modellek alkalmazása	63
4.2.7.	A modellek összehasonlítása	66
5.	Összefoglalás	71
6.	Köszönetnyilvánítás	73
7.	Irodalomjegyzék.....	74

1. Bevezetés

A gyógyszeripar jelenleg egyik legnagyobb kihívást jelentő feladata a Folyamatfelügyelő- és elemző Technológia (PAT) irányelveinek megvalósítása, mely megköveteli a gyógyszergyártóktól, hogy a gyártási folyamatokat megismerjék, nyomon kövessék és minél pontosabban szabályozzák. Ezáltal biztosíthatóvá válhat a termékek állandó, megfelelő minősége. Ennek megvalósítása olyan analitikai módszereket igényel, melyek mintaelőkészítés nélküli, roncsolásmentes mérésekkel tudnak megvalósítani valós vagy közel valós idejű vizsgálatokat a hatóanyag előállításától a végső minőségellenőrzésig. A Raman-térképezés ezeknek a célkitűzéseknek a realizálását segíti elő. A módszer alkalmazása PAT-ban már eddig is jelentős eredményeket hozott, azonban bizonyossá vált, hogy a térképek által szolgáltatott óriási mennyiségű adatból a kívánt információ kinyerésére a hagyományosan alkalmazott kiértékelési módszerek gyakran nem elegendők, így felmerül az igény kifinomultabb többváltozós adatelemzési technikák alkalmazására.

TDK munkám során két modellrendszer mennyiségi meghatározását végeztem el. Az egyik rendszer polimorfok keveréke volt. A gyógyszeriparban hangsúlyos ezek tisztaságának jellemzése, mivel polimorf szennyezők előfordulhatnak a hatóanyag kristályosításakor vagy akár más nem kívánatos mellékfolyamat eredményeként is. Ezek a szennyezések a termékek stabilitását jelentősen leronthatják. A másik tanulmányozott problémakör a hatóanyagot és egyéb segédanyagokat tartalmazó keverékek pontos kvantitatív meghatározása. Erre a készítménygyártás (például keverés, tablettázás) közben valamint a végső minőségellenőrzés, például a termék hatóanyagtartalom-egységességének („*content uniformity*”) vizsgálata során is szükség van. Az említett két rendszer pontos mennyiségi meghatározásával tehát szabadalmi, technológiai és minőségbiztosítási tekintetből is fontos tényezőkre derülhet fény.

Céлом az volt, hogy ezen gyógyszeripari illetve PAT szempontból is releváns rendszerek vizsgálatán keresztül tapasztalatot szerezzek arról, hogy milyen tényezők befolyásolják a mennyiségi meghatározás pontosságát egy kemometriai módszerekkel felépített modell esetén, különös tekintettel az eddig nem vagy csak ritkán használt technikákra. Ezek a tapasztalatok reményeim szerint hozzájárulhatnak a kvantitatív modellépítés sikeres megvalósításához, amely nélkülözhetetlen ahhoz, hogy a Raman-térképezés integrálása PAT-ba valamint a fellépő problémák megoldása realizálódjon.

Dolgozatomban bemutatom a két keverék rendszer kvantitatív meghatározására használt lineáris és nemlineáris módszereket, alkalmazásuk során felmerülő főbb kérdéseket és a nyert eredményeket összehasonlítom az általánosan használt Lambert-Beer törvény alapján végzett egyváltozós kiértékeléssel. Tanulmányozom különféle adatelőkészítési módszerek modelljóságra gyakorolt hatását valamint foglalkozok a spektrumokon végzett változó kiválasztási technikák vizsgálatával, melyek alkalmazására a Raman-térképezés területén eddig nem volt példa. Végül a modellek összehasonlítása után ismertetem azok új, a kalibráció és validálás során nem használt Raman-térképen történő alkalmazásának eredményeit.

2. Szakirodalmi áttekintés

2.1. A Raman-spektrometria kialakulása

A rezgési spektrometriai módszerek az atomok rezgési állapotának változásai alapján szolgáltatnak információt a vizsgált mintáról. Rezgési energia-átmenetek megvalósulhatnak, ha az anyag fénnyel (elektromágneses sugárzással) hat kölcsön, melynek három típusát különböztethetjük meg: az anyag a fény egy részét elnyeli (abszorpció), átereszt (transzmisszió) illetve bizonyos részét visszaveri (reflexió). Ezek aránya jellemző az anyagra, így beszélhetünk abszorpciós, transzmissziós és reflexiós színekéről.

A visszaverődés során a fotonok és a molekulák rugalmas ütközésének eredményeként a fény ugyanolyan energiával reflektálódik, mint amilyen a besugárzott fény energiája (Rayleigh-szórás). Egy részük azonban rugalmatlanul ütközik, amelynek következménye, hogy a foton a besugárzó frekvenciánál kisebb vagy nagyobb frekvenciával szóródik. Ez a Raman-spektrometria alapjául szolgáló Raman-szórás jelensége, melyet Brillouin [1] és Smekal [2] elméletben már az 1920-as években levezetett, majd ezt kísérleti úton Raman [3] bizonyította. A Raman-spektrum minőségi és mennyiségi analízisre egyaránt alkalmas. A minőségi (kvalitatív) információt – az infravörös spektrumokhoz hasonlóan – a spektrum ujjlenyomat jellege hordozza. A két spektroszkópai módszer kiválasztási szabályai azonban különböznek: míg az IR spektrumot a molekula dipólusmomentum-változása okozza, Raman-aktív az a vegyület lesz, melyben a polarizálhatósági tenzor változik. Az adott kötések sokszor csak az egyik spektroszkópai módszerre igazán érzékenyek, ezért a két vizsgálati módszer komplementer információt hordoz.

A mennyiségi (kvantitatív) analízis alapjául szolgáló 2.1. egyenletet 1934-ben Placzek állította fel polarizálhatósági elméletében [4]. Ez alapján egy adott hullámhosszúságú foton Raman-intenzitása a besugárzástól és az anyagi minőségtől a következő módon függ:

$$I(\nu)_R = \frac{2^3 \pi^3}{45 \cdot 3^2 \cdot c^4} * \frac{h I_L N (\nu_0 - \nu)^4}{\mu \nu \left(1 - e^{-\frac{h\nu}{kT}}\right)} * [45(\alpha'_a)^2 - 7(\gamma'_a)^2], \quad (2.1.)$$

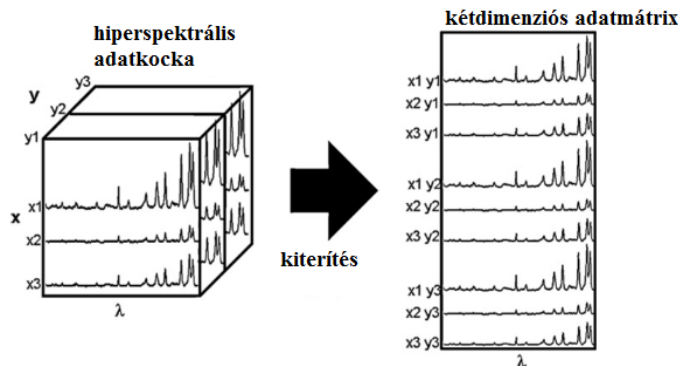
ahol c a fénysebesség, h a Planck-féle állandó, I_L a besugárzás intenzitása, N a fotonokkal ütköző molekulák száma, ν a molekula rezgési frekvenciája (Hz), ν_0 a besugárzó lézer frekvenciája (Hz), μ a rezgő atomok redukált tömege, k a Boltzmann-állandó, T az abszolút hőmérséklet, α'_a a polarizálhatósági tenzor izotróp tagjának nyoma, γ'_a a polarizálhatósági tenzor anizotróp tagjának második invariánsa. Az egyenletből megállapítható, hogy a Raman-intenzitás lineárisan függ a

fotonnal ütköző molekulák számától. Eszerint az adott hullámhosszon kapott jel arányos a jelet okozó kötések számával, azaz a vegyület koncentrációjával.

A besugárzott fotonoknak mindössze 10^{-6} - 10^{-8} -ad része eredményez Raman-fotonokat. Ennek egyrészt az a következménye, hogy köztük kölcsönhatás kis valószínűséggel lép fel és így keverékek összspektruma jó közelítéssel a komponensek spektrumainak lineáris szuperpozíciója. Másrészt a Raman-szórás intenzitásának gyengesége miatt a Raman-spektroszkópia széleskörű gyakorlati alkalmazása a lézerek felfedezéséig [5,6] nem volt megoldott. A 2.1. egyenlet alapján lézerrel mint nagy energiasűrűségű monokromatikus fénnyel való besugárzással a Raman-szórt fény intenzitása jelentősen növelhetővé vált, amely a mérési időt drasztikusan csökkentette. A műszerfejlesztések 1974-ben már oda vezettek, hogy térképezésre is alkalmas mikrospektrométer épült [7,8], a kémiai térképezés pedig jelenleg is a spektroszkópia legdinamikusabban fejlődő területei közé tartozik.

2.2. Kémiai térképezés és alkalmazásai a gyógyszeriparban

Kémiai térképezés során a spektrométert egy mikroszkóppal kapcsolva térbeli képalkotásra nyílik lehetőség, ezáltal a vizsgált mintáról a spektrális információ mellett térbeli információt is kapunk [9]. A Raman-térképezés megvalósításának bemutatásával több tanulmány is részletesen foglalkozik [9,10,11,12,13], melyekben három főbb mérési elrendezést különböztetnek meg. Az egyik módszer szerint mozgatható tárgyasztal segítségével történik a pontonkénti spektrumfelvétel (térképezés, „*point mapping*”). További lehetőségként egy vonal mentén egyszerre több spektrum is felvehető („*line mapping*”) illetve készülhet az egész mintafelületről is kép egyetlen hullámszámon (képalkotás, „*global imaging*”). A térképfelvétel eredményeként kapott adathalmaz „hiperspektrális adatkockába” rendeződik, melynek két éle a felvett pont koordinátáit, míg a harmadik a felvett hullámszámokat jelzi, így a kocka az egyes mérési pontokról felvett spektrumok adott hullámszámokon mért intenzitásértékeit tartalmazza. Az adatelemzés előtt ebből a struktúrából először kétdimenziós szerkezetet kell kialakítani. A kiterítést (*unfolding*) követően kétdimenziós mátrixot kapunk, mely az intenzitásértékeket az x-y helykoordináták és a λ hullámszámok függvényében tartalmazza (2.1. ábra).



2.1. ábra: A hiperspektrális adatkocka szerkezete és a kétdimenziós adatszerkezet kialakítása [14]

X és y a térkoordinátákat, λ a hullámszámokat jelöli

A térképezett területről akár több száz vagy ezer spektrum is készülhet, egy-egy spektrum pedig szintén száz-ezer hullámszámból állhat. A kapott mátrix ezért rengeteg adatot tartalmaz, amelyből a kívánt információ kinyerése nagy kihívást jelent. A spektrumok kiértékelése általában a Lambert-Beer törvényen alapszik, azonban ez az egyváltozós kiértékelés az adathalmaz összetettsége miatt nem minden esetben elegendő. Ezért merül fel a sokváltozós adatelemzési (kemometriai) módszerek alkalmazásának igénye, mely jobb kiértékelést tesz lehetővé és akár a mérési időt is csökkentheti. [15]

A kémiai térképezés népszerűségét jelzi, hogy 2000 és 2008 között a témával foglalkozó cikkek száma megduplázódott [10] és a technika jelentősége azóta is folyamatosan nő. Használják többek között az orvostudományban, például tumoros sejtek azonosítására [16] illetve élő sejtek vizsgálatára [17] és előfordul alkalmazása élelmiszerbiztonsági és -minőségi kérdésekben [18,19], a műanyagiparban [20] és bűnügyi kérdésekben is [21].

Különösen nagy szerepet kap a Raman-térképezés a gyógyszeriparban azáltal, hogy roncsolásmentes és gyors vizsgálattal nyújthat információt a gyógyszerkészítmények előállításának szinte bármely fokán, a hatóanyag előállításától a végső minőségellenőrzésig. Részletes összefoglaló tanulmányok mutatják be a NIR- és Raman-térképezéssel elért eredményeket [9,10,11]. Sacré [13] összefoglaló cikkében hét csoportra osztja a fő alkalmazási területeket: homogenitás- és részecskeméret-vizsgálatok, polimorfokkal kapcsolatos és oldódási vizsgálatok, hamisított gyógyszerek felderítése valamint a hatóanyag-tartalom egységességének biztosítására és a gyógyszeripari folyamatok megértésére irányuló kutatások. A cikk részletesen bemutatja a térképek értékeléséhez használt módszereket is.

A gyógyszeripar talán jelenlegi legnagyobb feladata az amerikai Élelmiszer- és Gyógyszer Ellenőrző Hatóság (FDA, *Food and Drug Administration*) 2004-ben meghirdetett irányelveinek való megfelelés. A Folyamatfelügyelő és Analizáló Technológia (PAT, *Process Analytical Technology*) [22] megköveteli a gyógyszergyártóktól, hogy a gyártásban történő folyamatokat megismerjék, szabályozzák, hogy ezzel biztosítva legyen a készítmények minőségének megfelelő, előre megadott határok között történő ingadozása. A folyamatok közbeni mérések és visszacsatolások a jelenlegi költség- és időigényes végellenőrzési lépéseket válthatják ki. A rezgési spektrometriai módszerek, köztük a Raman-spektrometria és a kémiai térképezés a gyártás folyamatához kiválóan illeszthetők, ezért kulcsszerepet játszhatnak a PAT megteremtésében [23]. Gowen [9] 2008-ban megjelent tanulmányában tekinti át ezeket a lehetőségeket, kitérve a megoldásra váró kérdésekre is. A műszerfejlesztés igénye mellett fő problémának az adatok kiértékelésére használt kemometriai módszerek hiányosságait említi. Ennek megoldására új adatelemzési módszerek alkalmazását javasolja, példaként említi a támogatóvektorok módszerét (SVM, *Support Vector Machine*). A kijelölt út ellenére Sacré már említett munkájában [13] hivatkozott publikációk közt csak kevés példa található ezen újszerű módszerek alkalmazására az általánosan használt kemometriai módszerek mellett. A két publikáció közt eltelt hat év eredményeinek ellenére ezen a területen tehát nem következett be lényeges fejlődés, habár a szerző ismét azt hangsúlyozza, hogy a kémiai térképezés területén bekövetkezett előrelépések új adatelemzési módszerek használatát követelik meg.

Egy, a PAT-ot kemometriai szemszögből vizsgáló cikk [24] a PAT-ot alkalmazó műveleteket és a használt adatelemzési módszereket tekinti át, melyből kitűnik, hogy az alkalmazott adatelemzési módszerek közel háromnegyede főkomponens-elemzés vagy -regresszió (PCA, PCR, *Principal Component Analysis/Regression*) és legkisebb négyzetek módszere (PLS, *Partial Least Squares*) [25]. E módszerek elméletét és használatuk fő kérdéseit Rajalahti [26] is részletesen tárgyalja. Az alkalmazott műveleteket tekintve, a PAT technológiának különös szerepe a hatóanyagokkal való műveleteknek és a kristályosításoknak van, valamint jelentős a keverés, szárítás és a tablettázás művelete [24]. Ez a felosztás arra is felhívja a figyelmünket, hogy ezekben a műveleti lépésekben fellépő kihívások azok, melyekre a kemometria eszköztárával kell választ találni.

Ezen feladatok közt említhetjük a kristályosítás illetve a hatóanyag előállítás során fellépő polimorfok pontos minőségi illetve mennyiségi analízisének igényét, valamint a hatóanyag-

segédanyag rendszerek kvantitatív meghatározását. Ezek számos technológiai és minőségi kérdésnél lehetnek kulcstényezők, amelyeknek lényeges szerepük van a PAT szempontjából. A továbbiakban ezeken a területeken tett vizsgálatokat tekintem át, különös tekintettel a Raman-térképezéssel elért eredményekre.

2.2.1. Polimorf keverékek mennyiségi meghatározása

A polimorfia az a jelenség, miszerint egy anyag különböző kristálmódosulatokban fordulhat elő. Polimorfia a szerves anyagok 80-90%-ánál figyelhető meg [27], így a gyógyszeripari gyakorlatban is különös jelentősége van, mert a különböző kristályszerkezet jelentős eltéréseket okozhat a fizikai-kémiai tulajdonságokban, például az olvadáspontban, oldhatóságban. Az ilyen szerkezetbeli különbségek hatással lehetnek akár az anyag stabilitására, biohasznosulására is. Ezen tényezők miatt a polimorfok vizsgálata kiemelten fontos kutatási terület. Szilárd fázisban erős mechanikai behatásra vagy hőhatásra a gyártás során is felléphetnek változások, ezért a megfelelő kvalitatív és kvantitatív analízis a műveletek során döntő fontosságú a végtermék minőségének biztosításához [28]. A téma fontosságát a polimorfok analízisével foglalkozó közlemények nagy száma is tükrözi: Chieng egyetlen összefoglaló munkájában [28] a szilárd-fázisú analízisre használt eszközök valamint a vizsgált gyógyszerhatóanyagok bemutatása közben több mint 200 cikkre hivatkozik.

Több kutató vizsgálta a Raman-spektrometria alkalmasságát polimorfok tanulmányozásában. Campbell [29] porröntgen-diffrakció és FT-Raman segítségével készült mérések összehasonlításáról számolt be, amelyeket mannitol polimorfok keverékein végzett el. A két módszer összevetésével azt állapította meg, hogy a csúcshintézetek arányán alapuló Raman-spektrometriás kiértékeléssel pontosabb meghatározás érhető el, mint a széles körben használt röntgen-diffrakcióval. Szintén erre az eredményre jutottak famotidin [30], piracetam [31] és paracetamol [32] polimorfok meghatározása során is. A mennyiségi becslésre, azaz a kvantitatív információ kinyerésére ezekben az esetekben már többváltozós adatelemzési módszereket használtak (PLS, PCR) valamint különböző adatelőkészítési módszerek (alapvonal-korrekción, multiplikatív szóródási korrekció (MSC, *multiplicative scatter correction*), sorstandardizálás (SNV *standard normal variate*)) hatását is tesztelték. A ranitidin-hidroklorid polimorfjainak vizsgálatára több példa is található. Chieng rámutatott [33], hogy Raman-spektrometriával két polimorf módosulat mellett az amorf hányad is megfelelően meghatározható, McGoverin pedig

megállapította, hogy a NIR-spektrometria pontossága kisebb a Raman-spektrometriához képest [34]. Munkatársaival szintén ő vizsgálta a transzmissziós Raman teljesítőképességét segédanyagok melletti, formulált mintákon [35]. Szulfamerazin polimorfok kvantifikálása kapcsán a Raman-térképezés, a DSC és a röntgendiffrakció összehasonlítását végezték el [36]. A három analitikai mérés közül a Raman térképek nyújtották a legkevesbé jó becslést az összetételre, azonban a leírt mérési paraméterek miatt (rossz nagyítás, kis lépésköz) nem kaphattak megfelelően reprezentatív adatokat a koncentrációra. Schönlichler és munkatársai [37] pedig furosemid három polimorfja segítségével elsőként hasonlították össze a NIR-kémiai térképezés hatékonyságát az azonos mintákon végzett NIR-, Raman- és az ATR-IR spektroszkópia teljesítőképességével. Megállapították, hogy a térképek spektrumaiból alkotott modellek használatával hasonló pontossággal becsülhető a polimorfok mennyisége, mint a hagyományos spektroszkópiai eljárásokkal. Azonban a NIR-térképek további információt adnak a komponensek eloszlására is. Hu és társai [38] hasonló megfontolásokból rezgési spektroszkópiai módszereket (ATR-IR, NIR, Raman) használtak szulfatiazol három polimorfból álló keverékének vizsgálatára. Kémiai térképeket ebben a közleményben nem írtak le, de ebben az esetben is tisztán megjelenik az a trend, hogy legtöbbször a PLS módszert alkalmazzák, ha a vizsgált polimorfok mennyiségére szeretnének következtetni.

Az indometacin polimorfok vizsgálatával kapcsolatban a kutatók stresszkörülmények közt fellépő változásokat kívántak nyomon követni [39,40]. Ugyanezzel a hatóanyaggal látta be Heinz [41], hogy a Raman-spektrometriai módszer még fluoreszcens háttér mellett is jó eredményeket ad. A tanulmányokban továbblépésként tekinthetjük a különböző kristálmódosulatú anyagok formulált mintákban való vizsgálatát. Ilyen kísérletsorozatot valósítottak meg piroxicam polimorfok NIR-térképezésével [42]. A két polimorf módosulatot laktóz, nátrium-hidrogénkarbonát, talkum, keményítő és Mg-sztearát segédanyagok mellett tablettázták. A PLS modellekkel nyert koncentráció-térképek alapján a két polimorf eltérő kristályformáit (tűs, gömbölyű) is azonosítani tudták. Hennigan egy tanulmányában [43] pedig a már több közleményben előfordult piracetam polimorfok kis koncentrációit vizsgálta a főtömegben kalcium-karbonátot és mikrokristályos cellulózt tartalmazó tablettákban. Vizsgálata során kitért a tabletták vastagságának hatására is, és arra, hogy transzmissziós Raman-spektroszkópiával végzett mérés szolgáltatta a legpontosabb eredményt.

A polimorfok PAT körülmények között való kvantitatív meghatározására két friss közlemény is található. Egyik esetben [44] karbamezapin polimorfok hő hatására történő átalakulását vizsgálták in-situ Raman-térképezéssel, majd a polimorfok eloszlását MCR-ALS módszerrel modellezték. Soares és társai [45] ibuprofen-nikotinamid kokristályok képzését követték nyomon több mérési módszerrel is. A porröntgen-diffrakció, DSC és ATR-FT-IR mellett a Raman-spektrometriával bizonyítottan megbízható eredményekhez jutottak.

Az eddig bemutatott eredményekből kitűnik, hogy a Raman-spektrometria a polimorfok mennyiségi meghatározásának egyik legígéretesebb eszköze. Az említett cikkekben azonban az adatelemzési módszerek nem lépnek túl a CLS- PCR- és PLS-regresszió, a pontosabb eredményekre főleg az adatelőkészítési módszerek változtatásával és kombinálásával (pl. alapvonal-korrektúra, normálás, SNV, MSC) törekednek a kutatók. Bár ezekkel is jelentős előrelépések érhetők el, érdemes lehet egyéb kemometriai módszereket is használni. Erre enged következtetni Okumura és munkatársainak [46] munkája is, akik porröntgen-diffrakció és mesterséges ideghálózatok segítségével végezték el az α - és γ -indometacin keverékekben a koncentrációk meghatározását. A meghatározás pontossága jóval felülmúlta a referenciamódszerként alkalmazott többváltozós lineáris regressziót. Az új kemometriai módszerek igényére Soares [45] is felhívja a figyelmet: a kokristályosítás eredményeihez intervallum-PLS változókiválasztási módszer segítségével jutottak.

2.2.2. Hatóanyag-segédanyag rendszerek kvantitatív meghatározása

A gyógyszerek nagy része szilárd formában kerül felhasználásra. A készítmények a hatóanyagok mellett különböző segédanyagokat is tartalmaznak, melyek egyrészt a kívánt hatás kiváltásért lehetnek felelősek, másrészt a készítmény megfelelő formájának előállításához szükségesek (pl. kötő- és csúsztató anyagok a tablettázás során). A készítmény végső tulajdonságait ezeknek a segédanyagoknak az eloszlása és mennyisége döntően befolyásolja, ezért nem hagyhatjuk figyelmen kívül a szerepüket. A formálási lépésekben létrejövő keverékek összetétele, a komponensek eloszlása jellemzi a gyártási folyamat minőségét, amely kihat a készítmény végső minőségére. Ennek megfelelően a PAT fő feladatai közé tartozik a hatóanyag-segédanyag keverékek összetételének megismerése, állandó követése és szabályozása. E kérdéssel találkozhatunk a gyógyszergyártás folyamata során a keverékek előállításakor vagy tablettázáskor illetve a végső készítmény hatóanyagtartalom-egységességének (*content*

uniformity) vizsgálatakor is [23]. Nem véletlen tehát, hogy ebben a témában megjelent cikkek jelentős része a hatóanyag-segédanyag rendszerek kvantitatív meghatározásával a PAT kontextusában foglalkozik.

Ravn [47] NIR-térképezésen keresztül mutat be szilárd gyógyszerformák mennyiségi meghatározására alkalmas módszereket, az egy kiválasztott hullámszám segítségével történő kalibrálást, a klasszikus legkisebb négyzetek módszerét (CLS, *classical least squares*) és a PLS modellezést. A hatóanyag mellett cellulózt, laktózt, magnézium-sztearátot és talkumot tartalmazó tabletták analízise során arra a következtésre jutott, hogy a meghatározás pontosabbá tehető kalibrációs sorok használatával, mintha csak a komponensek tiszta spektrumait használó modellekből nyernénk ki a koncentrációkat. Egy másik közleményben [48] a többváltozós lineáris regressziót és a PLS módszert mikronizált hatóanyag és segédanyagok in-line mennyiségi meghatározására használták.

Az összetevők homogenitásának vizsgálatára kitűnő módszerként adódik a kémiai térképezés módszere: a térképek koncentráció-eloszlásából a minta homogenitására különböző indexek használatával következtethetünk [49,50]. Ez a módszer nyilvánvalóan csak akkor szolgáltathatja a valós eredményeket, ha a koncentrációk meghatározása megfelelő pontossággal történik. Gendrin és munkatársai [51] a hatóanyag-eloszlás egységességét vizsgálták hétkomponensű tablettában. A hatóanyag és két segédanyag koncentrációit változtatták, majd CLS- és PLS-modellekkel és különböző adatelőkészítéssel hozták létre az összetevők koncentráció-térképeit. Tapasztalatuk szerint a csak tiszta komponenseket használó CLS-modellek a hatóanyagot tekintve kevésbé érzékenyek a kiugró adatokra és jobban korreláltak a valós összetétellel. Azonban a cellulóz és a laktóz mennyiségi elemzése során már problémákba ütközött a módszer alkalmazása. Hasonló munkáról számoltak be Sabin és munkatársai, közleményükben különböző kiértékelési módszerekkel előállított eloszlástérképeket hasonlítottak össze [52]. A carbamazepin hatóanyagot és hat segédanyagot tartalmazó keverékeket többváltozós lineáris regresszió (MLR), többváltozós görbefeletbontás (MCR) valamint CLS és PLS regresszió (PLS) vetették alá és megállapították, hogy az utóbbi három módszer hasonló eloszlástérképeket szolgáltat, míg az MLR nem ad megbízható eredményt. További példa CLS modellezés alkalmazására [53], hol különböző technológiákkal készült tabletták jellemzését végezték el, illetve hasonlították össze az eredményeket SEM/EDX felvételekkel. Példa a keverés műveletének in-line követésére és homogenitás vizsgálatára is [54], mely olyan esetet tárgyal,

ahol a hatóanyag és három segédanyag összekeverésének folyamatát követték nyomon NIR-spektrometria segítségével olyan módon, hogy offline felállított PLS-modellt használtak a valós idejű detektálás során. Bár az eredmények alapján ez is járható útnak tűnik, a PLS- modell in-line módon történő felállítása [55] további előrelépést jelent.

Az utóbbi néhány évben ígéretes törekvések történtek a hatóanyagok és segédanyagok pontos analizisére bliszteren keresztül [56], a tabletta bevonás területén [57] vagy éppen több technológiai lépésből álló folyamat nyomon követésére [58] is. A kísérletek megvalósításakor szinte kizárólagosan a PLS- modellezés volt a fő kemometriai eszköz. Nem vizsgálták egyik munkában sem, hogy milyen egyéb kemometriai módszereket lehetne alkalmazni a folyamatok megismerésére felhasznált modellek további javítására. Fontos szempont, hogy milyen modelleket építsünk, ha azokat valós idejű detektálásra szeretnénk használni. Ennek főbb kérdéseivel Nakagawa [59] foglalkozik közleményében, melyben kiemeli, hogy a pontosabb modellek érdekében érdemes lehet változókiválasztási módszereket használni. Javasolja a helyileg súlyozott PLS modellek vizsgálatát is, mellyel az általa használt keverékrendszerben a legpontosabb kvantifikálást érte el. Ennek a módszernek használatára még egy példa található, hatkomponensű keverék előállítása során használták a hatóanyag mennyiségi meghatározására [60].

Egyéb kemometriai módszerek használatára is találhatók példák a gyógyszeriparban, ígéretességük ellenére a polimorfok és a hatóanyag-segédanyag rendszerek mennyiségi meghatározására még nem vagy csak igen kevés kísérlet történt ezekkel. A továbbiakban ezeket a módszereket kívánom bemutatni.

2.3. Ritkábban alkalmazott kemometriai módszerek a mennyiségi vizsgálatokban

Kvantitatív modelljeink pontosabbá tételének egyik útja a változókiválasztási módszerek alkalmazása lehet. Egy spektrum digitalizált formában több száz változót (hullámszámot) tartalmazhat, de ezek közül nem mindegyik hordoz lényegi információt a mintában lévő komponensekről. Bár a főkomponens-analízisen alapuló módszerek esetén is dimenziócsökkentés történik, a főkomponensek megalkotása az összes változó figyelembevételével történik, ennek okán a nem informatív változók is kifejtik hatásukat [61]. Ezért célszerű, hogy ezeket a változókat a modellépítés előtt eltávolítsuk az adathalmazból. Ehhez több lehetőség is rendelkezésre áll, melyeket Mehmood gyűjtött össze [62]. Tanulmányában felhívja arra a

figyelmet, hogy a módszerek teljesítőképessége az adathalmaztól függ, így a megfelelő eredmények több próbát is igényelhetnek.

Talán az egyik leggyakrabban alkalmazott változókiválasztási módszer a genetikus algoritmusokon alapuló módszer [63,64]. Gyógyszeripari alkalmazására is sor került már vitamintabletták B₁₂ tartalmának meghatározásakor [65], összehasonlítása az intervallum-PLS-el [66] pedig Leardi és Norgaard közös közleményében található [67]. A genetikus algoritmusok többváltozós kalibrációs módszerekkel való alkalmazásának lehetőségeivel foglalkoznak kínai kutatók [68]. Az intervallum- PLS alkalmazására a NIR spektroszkópia területén találunk példát [69], polimorfokkal kapcsolatos használatára pedig említhetjük a már bemutatott közleményt [45]. Más változókiválasztási módszert már használtak korábban élelmiszeripari minták NIR-térképezésében [70].

A kemometriai módszerek másik nagy jelentőségű csoportjába a nemlineáris modellezési lehetőségek tartoznak. A helyileg súlyozott regresszió (LWR) [71] használatára már bemutatott példák mellett [59,60] kevés egyéb alkalmazást találhatunk. Pérez-Marin a nemlineáris módszereket NIR-spektroszkópiái szemszögből tárgyaló közleményében [72] is csupán egy gyógyszeripari alkalmazás található.

A mesterséges ideghálók [72] jó módszernek bizonyultak kis illetve nagy nemlinearitás kezelésére is a NIR-spektroszkópiái használata során [73]. Alkalmazására több példa található az UV-VIS spektroszkópia területén gyógyszerkészítmények és hatóanyag-keverékek vizsgálatára [74,75,76,77]. Ezekben az alkalmazásokban különösen jó módszernek akkor bizonyult, amikor a genetikus algoritmussal végzett kalibrációt változókiválasztással kombinálták. Raman-spektroszkópiával eddig tabletták atorvasztatin-kalcium tartalmának meghatározására használták [78], az élelmiszeriparban pedig cukortartalom vizsgálata történt a módszer segítségével [79]. Polimorfok FT-Raman-spektrometriával történő elemzésére is található példa a szakirodalomban [80].

A támogatóvektorok módszerének (SVR, *support vector regression*) [81,82] elterjedésében sem történtek jelentősebb előrelépések. Thiessen [83] kisebb felbontású Raman-spektrométerrel felvett spektrumok segítségével is jó meghatározást tudott elérni a módszer segítségével, viszont arra is felhívja a figyelmet, hogy a módszer használata a viszonylag lassú kalibráció miatt nem mindig javasolt. Egyéb nemlineáris módszerekkel való összehasonlításából [73] azt szűrték le, hogy hasonlóan jó eredmények érhetők el vele, mint a mesterséges

ideghálókka, ellenben a számítási igény nagy kalibráló adathalmaz esetén megnőhet. A tanulmány szerint azonban a spektroszkópiai alkalmazások esetén ez általában nem számottevő probléma. A módszert a gyógyszeriparban UV-spektrumokon alkalmazták két hatóanyag szennyezők jelenlétében történő meghatározására [84].

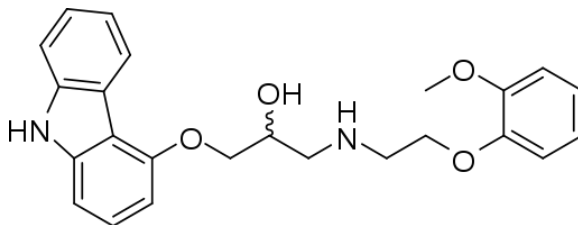
A fellelhető szakirodalom áttekintése rávilágít arra, hogy a gyógyszeripar egyre kifinomultabb kemometriai módszerek alkalmazását igényli, különösen a Folyamatfelügyelő és Analizáló Technológia (PAT) megvalósításához. Két különösen fontos szegmens a polimorf illetve a hatóanyag- segédanyag keverékek kémiai térképezéssel történő pontos kvantitatív meghatározása, mely téren eddig nem történt jelentős előrelépés. Munkám során ezért az volt a célom, hogy ezen rendszerek Raman-térképezéssel történő mennyiségi analízise során felderítsem a kiértékelés során alkalmazott kemometriai módszerekkel elérhető meghatározás pontosságát, valamint a használatuk során fellépő főbb kérdésekről, problémákról is tapasztalatot szerezzünk.

3. Felhasznált anyagok és módszerek

3.1. A kísérletek során felhasznált anyagok és mintaelőkészítési módszerek

3.1.1. Polimorf rendszer mennyiségi elemzése

A vizsgálatokhoz modellvegyületként carvedilol hatóanyag (3.1.1. ábra) Form I és Form II polimorf módosulatát választottam, mely nem-szelektív béta-blokkoló hatású szívgyógyszer.



3.1.1. ábra: Carvedilol hatóanyag szerkezete ($M = 406,5$ g/mol)

A kereskedelmi forgalomban kapható kinetikailag stabil II-es polimorfból a termodinamikailag stabil I-es formát oltásos átkristályosítással állítottam elő a következő módon. A carvedilol Form II módosulatából kimértem 25 grammot és egy háromnyakú gömblombikban, 120 ml etil-acetátban szuszpendáltam. A hőmérővel és hűtővel ellátott lombikot 100 1/perc fordulatszám mellett olajfürdőben 50 °C-ig fűtöttem. A kívánt hőmérséklet elérésekor beoltottam előzetes kísérletekből származó Form I módosulat 2,5 grammjával. Állandó keverés mellett 50 °C-on tartottam az elegyet 3 órán keresztül. A fűtés és keverés leállítása után az elegyet szobahőmérsékletig hagytam hűlni, majd üvegszűrőre öntve, vákuum és szívópalack segítségével szárazra szívattam. Raman-térképezés és porröntgen-diffrakció segítségével állapítottam meg, hogy az átkristályosodás lezajlott, Form II szennyezőt nem volt detektálható a Form I módosulat mellett.

A mennyiségi meghatározáshoz a polimorf módosulatokat különböző koncentrációkban tartalmazó, 19 keverékből álló mintasorozatot készítettem. A polimorfokat megfelelő mennyiségben összemértem, majd dörzsmozsárban homogenizáltam a keveréket. Erre azért is volt szükség, hogy a polimorfok egymástól eltérő alakú és méretű kristályszemcséit a jobb elegyíthetőség és a fajtázódás elkerülése érdekében közel azonos méretűvé őröljük. A 3.1.1. táblázat tartalmazza az egyes minták hivatkozási neveit, százalékos összetételeit valamint kalibráló és validáló halmazba való osztását. Ez utóbbi szerepe a 3.4. pontban kerül bemutatásra.

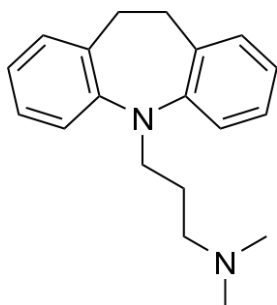
3.1.1. táblázat: A vizsgált Carvedilol polimorf keverékek százalékos összetétele és felosztása kalibráló és validáló halmazokra

Minta neve	Form I	Form II	halmaz típusa	Minta neve	Form I	Form II	halmaz típusa
CVD_1	0%	100%	kal.	CVD_11	70%	30%	kal.
CVD_2	1%	99%	kal.	CVD_12	85%	15%	kal.
CVD_3	2%	98%	val.	CVD_13	89%	11%	val.
CVD_4	3%	97%	kal.	CVD_14	92%	8%	kal.
CVD_5	5%	95%	val.	CVD_15	95%	5%	val.
CVD_6	8%	92%	kal.	CVD_16	97%	3%	kal.
CVD_7	11%	89%	val.	CVD_17	98%	2%	val.
CVD_8	15%	85%	kal.	CVD_18	99%	1%	kal.
CVD_9	30%	70%	kal.	CVD_19	100%	0%	kal.
CVD_10	50%	50%	val.				

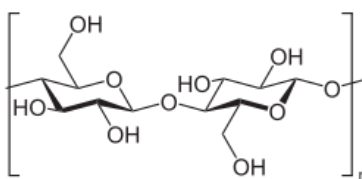
3.1.2. Hatóanyagot és segédanyagokat tartalmazó keverékek mennyiségi elemzése

A hatóanyag-segédanyag rendszerek mennyiségi meghatározásához olyan háromkomponensű keveréket alkalmaztam, melyben a modell-hatóanyag az imipramin (triciklikus antidepresszáns) volt (3.1.2.1. ábra), segédanyagként pedig cellulózt és keményítőt tartalmazott, melyeket széles körben használnak a gyógyszerkészítmények formulálása során. A keményítő és cellulóz molekulaszervezete nagyon hasonló: a cellulóz 1-4 kötéssel összekapcsolt β -D-glükopiranoz egységeket tartalmaz (3.1.2.2. ábra), a keményítő pedig α -D-glükopiranoz egységekből álló makromolekula, mely az 1-4 kötéseken kívül tartalmazhat 1-6 kötéseket is. A keményítőben az így kialakuló helikális szerkezetű rész az amilóz, az elágazásokat tartalmazót pedig amilopektinnek nevezzük. (3.1.2.3. ábra)

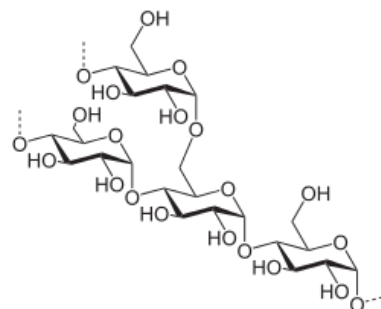
A mintasorozat készítése során mindhárom komponens esetén azonos mérettartományú szemcséket kevertem össze, melyet szitálással biztosítottam. Ezáltal kiküszöbölhető az a jelenség, hogy a mennyiségi meghatározást befolyásolja az összetevők eltérő szemcsemérete és az esetleges fajtázódás. A keverékek hivatkozási neveit és összetételét a 3.1.2. táblázat tartalmazza.



3.1.2.1.ábra: Imipramin
szerkezete (M=280,407 g/mol)



3.1.2.2.ábra: Cellulóz
szerkezete



3.1.2.3. ábra: Keményítő
(amilopektin) szerkezete

3.1.2. táblázat: A vizsgált hatóanyag-segédanyag keverékek százalékos összetétele

Minta neve	Keményítő	Cellulóz	Imipramin
M_1	0 %	0 %	100 %
M_2	0 %	25 %	75 %
M_3	25 %	0 %	75 %
M_4	0 %	50 %	50 %
M_5	25 %	25 %	50 %
M_6	50 %	0 %	50 %
M_7	0 %	75 %	25 %
M_8	25 %	50 %	25 %
M_9	50 %	25 %	25 %
M_10	75 %	0 %	25 %

Minta neve	Keményítő	Cellulóz	Imipramin
M_11	0 %	100 %	0 %
M_12	25 %	75 %	0 %
M_13	50 %	50 %	0 %
M_14	75 %	25 %	0 %
M_15	100 %	0 %	0 %
M_16	75 %	12,5 %	12,5 %
M_17	12,5 %	75 %	12,5 %
M_18	12,5 %	12,5 %	75,0 %
M_19	33,3 %	33,3 %	33,4 %

3.2. A kísérletek során használt berendezések és vizsgálati körülmények

3.2.1. A berendezések adatai

Röntgen-diffrakció:

A polimorf keverékek porröntgen-diffrakciós vizsgálatait X'pert Pro MPRD típusú PANalytical készülékkel készítették. A műszer 128 csatornás, 2°-ot mér egyszerre és 15 mm széles röntgennyalábbal végzi a vizsgálatot. A mérés során a röntgencső feszültsége 40 kV, áramerőssége 30 mA volt. A mintát Cu K α sugárással vizsgáltuk ($\lambda=1,5406 \text{ \AA}$) Kb (Ni) szűrő alkalmazása mellett. A mérési tartomány $2\Theta=2^\circ\text{--}42^\circ$ volt.

Raman-mikrospektrométer:

A Raman vizsgálatokat a Jobin Yvon cég által gyártott Labram típusú Raman mikrospektrométeren végeztem. A minta gerjesztéséhez háromféle lézer sugárforrás állt rendelkezésre, ezek közül a He-Ne gázlézert nem használtam, az alkalmazott Nd-YAG és diódalézer tulajdonságait a 3.2.1. táblázatban foglalom össze.

A vizsgálatokhoz alkalmazott lézer kiválasztásakor figyelembe kell venni, hogy bár a frekvenciakettőzött Nd-YAG lézerrel rövid mérési idő esetén is jó jel/zaj viszony érhető el, használatát a látható tartományban sokszor jelentkező erős fluoreszcens háttér gátolhatja, illetve a nagy teljesítményének hatására degradálódhat a minta. Ilyen esetben azonban sikerrel lehet alkalmazható a diódalézer.

A minta roncsolódásának elkerüléseért semleges szűrőket is a gerjesztő sugár útjába állíthatunk, melyek a lézerfény intenzitását csökkentik 1/2, 1/4, 1/10, 1/100, 1/1000 vagy 1/10000-ed részére valamint látható hullámhosszúságú lézer esetén depolarizátor is helyezhető a fény útjába.

A lézersugár egy megfelelő optikai tükör-és szűrőrendszeren jut az Olympus BXC-40 típusú optikai mikroszkópba, mely mikrométeres felbontást biztosít, majd cserélhető objektíveken keresztül éri el a mintát. Az objektívek 10×, 50× és 100× nagyítást tesznek lehetővé. Rendelkezésre áll egy makrolencse is nagy mintafelszín vizsgálatára, illetve fluidumok esetén az objektívek helyére küvetta szerelhető. A mintán szóródó fotonokat szintén az objektívek gyűjtik össze és juttatják a sugárzást a gerjesztő frekvenciának megfelelő „notch-szűrőkre”, mely a besugárzott fény frekvenciájával azonos frekvenciájú Rayleigh-fotonokat távolítja el. A Raman-sugárzás felbontására ezután kétféle, 1800 és 950 vonás/mm vonalsűrűségű diffrakciós rácsot iktathatunk be. Ezáltal körülbelül 3 cm^{-1} spektrális felbontás érhető el.

A Raman-mikroszkóp térbeli felbontása a fókuszpontban gerjesztődő mintatérfogattól függ, amit elsősorban az alkalmazott objektív befolyásol. Míg a nagyítás mértéke a laterális felbontásra van hatással, a numerikus apertúra értéke, mely az objektíven áthaladó fény törésszögét jellemzi, a mélységi felbontásra. A lézernyaláb fókuszálása ellenére a minta a fókuszsíktól eltérő síkban is gerjesztődik. Az innen származó Raman-fotonok kiszűrésére alkalmas a konfokális elrendezés, mellyel a nem kívánt mélységből érkező fotonokat a detektor előtt egy diafragma (kis átmérőjű lyuk) segítségével lehet kiszűrni azáltal, hogy a diffrakciós rács előtt a fénysugarat egy gyűjtőlencsén vezetik át, melynek gyújtópontja a diafragma közepére

esik. Így az a foton, mely nem a kívánt, lézersugárra merőleges mélységből érkezik, nem jut át a detektorra, ezáltal javítva a mélységélességet.

3.2.1. táblázat: A használt lézerek és objektívek jellemző paraméterei (*„NA”: numerikus apertúra)

		Frekvenciakettőzött Nd-YAG	Dióda
Hullámhossz [nm]		532	785
Energia [eV]		2,33	1,58
Teljesítmény a mintán [mW]		15	19
A lézersugár számított átmérője a fókuszpontban az egyes objektívek esetén [μm]	10×; NA *=0,25	2,553	3,768
	50×; NA *= 0,7	0,912	1,346
	100×; NA* = 0,9	0,709	1,047

A Raman-spektrum a Raman-eltolódás függvényében a detektált Raman-fotonok számával arányos intenzitásértékeket tartalmazza arányos egységben (a.u. „*arbitrary unit*”). A fotonok érzékelése 256×1024 pixeles CCD (*Charge-coupled device*) detektorral történik, mely beütésszáma maximált, tehát a pixelek telítődése esetén azok kiolvasásáig nem alkalmasak további detektálásra. Ezért jelent problémát a nagy fluoreszcens háttér jelenléte is, mert ekkor nem tudunk megfelelő jel/zaj arányt elérni még azelőtt, hogy telítenénk a detektorunkat. A telítődés az intenzitásskála 32 000-es értékénél történik meg.

A bemutatott módon összeállított berendezés motorizált tárgyasztal segítségével képes egy meghatározott mintafelület pontjainak program szerinti spektrumfelvételére, ezáltal kémiai térkép létrehozására. A minta felületének egyenetlensége azonban okozhat problémát a fókuszálás során. Ez egy automatikus fókuszáló rendszer segítségével oldható meg: az autofókusz a mintafelületről visszavert fény intenzitása alapján a fókuszpontot az objektív piezo-kristály segítségével történő mozgatásával módosítja. Azonban ennek használatakor azt is figyelembe kell vennünk, hogy jelentősen megnöveli a mérés összidejét.

A készüléket mérés előtt (naponta egyszer) kalibrálni kell, melyet szilícium lap segítségével végzünk. Ehhez a rácsot a lézer kikapcsolt állapotában fizikai nullára állítjuk, és a szűrőn beszűrődő fotonok általi csúcsmaximumot 0,0 cm⁻¹-re állítjuk a „Zero” érték változtatásával. A szilícium Raman-spektruma egyetlen, 520,5 cm⁻¹-nél detektálható csúcsból áll. A lapra 100× nagyítású objektívvel fókuszálva, majd a fókuszálást folytatva, 850 cm⁻¹ központú

spektrumot veszünk fel fél másodpercenként történő valós idejű spektrumfelvétellel egészen addig, míg el nem érjük a csúcs intenzitásának maximumát. Erre a sávra Gauss-Lorentz függvényt illesztünk, majd ha a függvény maximuma nem a kívánt $520,5\text{ cm}^{-1}$ hullámszámon jelentkezik, a koefficiens értékét állítjuk úgy, hogy a maximum erre a hullámszámmra essen. Mivel a rács relatív pozíciója változik ekkor, az optikai rács állását ismét 850 cm^{-1} -re kell állítani. Ismételt spektrumfelvétellel és függvényillesztéssel addig iterálunk a bemutatott módon, míg el nem éri a spektrum csúcsának helyzete a pontos $520,5\text{ cm}^{-1}$ -t.

3.2.2. A minták vizsgálati körülményei

3.2.2.1. Polimorfrendszer mennyiségi elemzése

Az előállított polimorf keverékek Raman-térképezéséhez az 532 nm -es Nd-YAG frekvenciakettőzött lézert használtam. Mindegyik térkép $10\times$ -es optikai nagyítással, intenzitásszűrő használata nélkül készült. A felvett spektrális tartományt $345\text{--}1790\text{ cm}^{-1}$ -re állítottam be. A spektrumok 1 s spektrumfelvételi idővel és 3 akkumulációval készültek (mely a véletlen zajok kiszűrését tette lehetővé). Ilyen beállításokkal a spektrumok intenzitása valamint a jel-zaj viszony is megfelelőnek adódott az alkalmazott hatóanyagmolekula erős Raman-aktivitásának és a minta nem túl erős fluoreszcenciájának köszönhetően.

Minden mintáról $441\text{ (}21\times 21\text{)}$ képpontból álló Raman-térkép készült. A térképezéshez a megfelelően homogenizált keverékekből kis mennyiséget tárgylemezre helyeztem, majd spatulával a felületét elegyengettem, hogy a térképfelvételt fókuszálási problémák ne befolyásolják. Egyéb minta-előkészítésre nem volt szükség.

A képpontok közötti lépésközt $50\text{ }\mu\text{m}$ -nek állítottam, így a feltérképezett terület 1 mm^2 -nek adódott. Ezzel a felvételi technikával egyrészt biztosítottam, hogy a felvett pontok függetlenek legyenek egymástól, másrészt a térképezett terület nagyságának köszönhetően a fizikai keverékben előforduló inhomogenitás a mennyiségi meghatározást ne befolyásolja. Ennek további biztosítása érdekében minden keverékről 3 különböző helyen készítettem térképet. Munkám során megvizsgáltam, hogy ez a térképméret valóban reprezentatív-e, azaz megfelelően homogén-e a felvett térkép. Ezt a makropixel-elemzés módszerével végeztem, melyet a 3.7. pontban mutatok be.

A mennyiségi meghatározás során épített modellek előrejelző képességének vizsgálatához az 5% Form I polimorfot tartalmazó keverékről a leírt paraméterekkel felvettem egy jóval nagyobb, 10201 (101×101) pontból álló térképet is.

3.2.2.2. Hatóanyagot és segédanyagokat tartalmazó keverékek mennyiségi elemzése

Az előállított hatóanyag-segédanyag rendszerben jelentős fluoreszcenciát tapasztaltam, ezért mérését a 784,08 nm-es diódalézerrel valósítottam meg. A felvett spektrális tartomány 458-1678 cm^{-1} , a mérésekhez 10 $\times$ -es objektívet használtam.

A megfelelő jel-zaj arány illetve a térképek felvételi idejének figyelembevétele alapján egy spektrum 30 másodperces spektrumfelvételi idővel, 2 akkumulációval készült. A keverékek tárgylapra helyezett, spatulával elegyengetett mintáiról 169 (13×13) pontból készült térképet vettem fel, a pontok között 50 μm -es lépésközzel. Az optimális térképméretet makropixel-számítás segítségével határoztam meg. Ehhez szükség volt egy nagyobb térképre is: az M9-es jelű (50% keményítő, 25% cellulóz és 25% imipramin tartalmú) mintáról vettem fel 2401 (49×49) pontból álló térképet.

3.3. A számításokhoz alkalmazott szoftverek

A spektrumok felvétele Labspec 5.41 (Horiba) programmal történt, melyben lehetőség van a spektrumok és térképek alapvető kiértékelésére is (alapvonal-korrekción, normálás, referenciaspektrumokkal végzett CLS modellezés stb.). Szintén ezzel a programmal végeztem a modellek által számolt koncentrációtérképek ábrázolását és vizuális összehasonlítását.

A kemometriai számításokat illetve adatelőkészítési módszereket Matlab 8.2. (MathWorks) programmal, valamint ennek PLS Toolbox 7.8.2. (Eigenvector Research) és Neural Network Toolbox 8.0.1. (MathWorks) bővítményeivel végeztem. Az egyváltozós és statisztikai számításokat Excel 2010 (Microsoft) és Statistica 8.0 (StatSoft) programmal végeztem, az SRD számításokhoz pedig a módszer kifejlesztői által rendelkezésünkre bocsátott Excelben írt VBA makrót használtam.

3.4. A modellek validálására használt módszerek és a modelljóságot jellemző paraméterek

Célom az volt, hogy a kemometriai módszerekkel készült modellek az építéshez használt minták jellemzésén túl alkalmasak legyenek jövőbeni minták megfelelő leírására is, azaz legyen egyfajta robosztusságuk, ne legyenek túlillesztettek. Ahhoz hogy a modellek pontosságát megismerjük, azok validálására van szükség. Ennek érdekében a mintasorozatokot felosztottam kalibráló és validáló halmazra. A modellek felépítéséhez illetve a modellparaméterek optimalására kizárólag a kalibráló halmazba tartozó mintákat használtam fel. Ezáltal a validáló halmaz elemei teljesen függetlenek voltak a modellektől, amely biztosította azt, hogy jól használhatók lesznek az előrebecslő képesség jellemzésére. (Az ilyen vizsgálatot a modellek *külső validálásának* nevezzük.)

A túlillesztés elkerülése érdekében belső validálást, azaz *keresztellenőrzést* végeztem, mely során a kalibráló halmazból egy-egy koncentrációsintet kizárva építettem fel a modelleket, majd azokat a kihagyott minták becslésére használtam. (A túlillesztés az a jelenség, miszerint a modell a kalibráló mintákban lévő lényegtelenebb információra és a zajokra is illeszkedik, ezáltal predikciós erejét elveszti.) Az összes minta ily módon becsült értékeiből számoltam ki a keresztellenőrzést jellemző paramétereket.

Meghatározás közepes négyzetes hibája (RMSE):

A modellek összehasonlítására szolgáló egyik legfontosabb paraméter a meghatározás közepes négyzetes hibája, melyet a következő képlet alapján számolunk:

$$RMSE = \sqrt{\frac{\sum_{i=1}^N \left(y_i^{(\text{számított})} - y_i^{(\text{valódi})} \right)^2}{N}}, \quad (3.1.)$$

ahol N a számítás során figyelembe vett minták száma, $y_i^{(\text{számított})}$ a modell által becsült koncentrációérték, $y_i^{(\text{valódi})}$ pedig a bemért koncentráció. Ez az érték külön számítható a kalibráló (RMSEC, *root mean square error of calibration*) validáló (RMSEP, *root mean square error of prediction*) halmazra és a keresztellenőrzésre is (RMSECV, *root mean square error of cross-validation*). A modellek túlillesztettségét jól mutatja a keresztellenőrzési hiba nagysága, illetve ennek és a predikciós hibának a különbsége is. Túlillesztés esetén ugyanis a keresztellenőrzési hiba jóval nagyobb a kalibráció hibájánál, és az előrebecslés közepes négyzetes hibája is jelentősen nagyobbá válik, mint a keresztellenőrzésé.

Determinációs együttható (R^2):

A modellek illeszkedését jellemezhetjük a determinációs együtthatóval, mely megmutatja, hogy a becslés a mért y_i értékek varianciájából mennyit magyaráz meg, azaz milyen mértékben írja le a mérendő változó viselkedését. Értéke 0 és 1 közötti, annál nagyobb, minél jobb a regresszió.

$$R^2 = 1 - \frac{\sum_{i=1}^N \left(y_i^{(\text{számított})} - y_i^{(\text{valódi})} \right)^2}{\sum_{i=1}^N \left(y_i^{(\text{valódi})} - \bar{y}^{(\text{valódi})} \right)^2}, \quad (3.2.)$$

$$\text{ahol } \bar{y}^{(\text{valódi})} = \sum_{i=1}^N \frac{y_i^{(\text{valódi})}}{N} \quad (3.3.)$$

Ezt az értéket szintén külön számíthatjuk a kalibráló (R^2_{cal}), validáló halmazra (R^2_{val}) és a keresztellenőrzésre (R^2_{CV}) is.

Tengelymetszet (Bias):

A regressziós egyenes ideális esetben 0 tengelymetszetű, 1 meredekségű egyenes, az ettől való eltérés a regresszió jóságának jellemzésére használható. Szintén külön számolható a kalibráló (Bias_{cal}) és validáló halmazra (Bias_{val}) és a tengelymetszetre (Bias_{CV}) is. Ez a modelljószág jellemző paraméter azonban kevésbé informatív, mint a meghatározás hibája és a determináció együttható, ezért a modellek optimalizálása során kisebb szerepet kapott.

Reziduumok:

A reziduumok a becslt és a bemért koncentrációk különbségei, melyet a bemért koncentrációk függvényében ábrázoltam minden modell esetén. Az így kapott ábra segítségével érzékenyen vizsgálható a modell adekvátsága.

Rangszámkülönbségek abszolútérték-összegének módszere, SRD [85]:

A már létrehozott modellek összehasonlítására a Rangszámkülönbségek abszolútérték-összegének módszerét (SRD, *Sum of ranking differences*) is alkalmaztam, mely egy nemparaméteres statisztikai módszer.

A módszer alkalmas arra, hogy modellek, módszerek jóságát tudjuk vizsgálni egy meghatározott referenciához képest. A referencia lehet egy általunk megadott adatsor (például becslt koncentrációk rangsorolása esetén a bemért koncentrációk), de használhatjuk a rangsorolni kívánt adatok minimumát, maximumát illetve átlagát is. Ezáltal lehetőség van különböző szempontok szerint jellemezni az összehasonlítani kívánt adatsort. Az SRD

megvizsgálja az egyes esetek eltérésének nagyságát a referenciától és ezek abszolút értékét minden rangsorolni kívánt változó esetén összegzi. A referenciára leginkább hasonlító modell adja a legkisebb rangszámösszeg értéket, ez alapján a modellek sorrendbe állíthatók.

Annak vizsgálatára, hogy ez a rangsor a véletlen műve-e, a véletlen számok rangsorolásával nyert Gauss-görbét vehetjük figyelembe. A rangsorolás akkor tekinthető jónak, ha a görbétől a rangszámösszegek elkülönülnek.

3.5. Alkalmazott adatelőkészítési módszerek

A felvett térképek háromdimenziós, „hiperspektrális adattömb” struktúrájúak ($n \times m \times \lambda$, n és m a pontok száma a X és az Y irányban, λ a hullámszám-csatornák száma). Ennek kétdimenzióssá történő kiterítését („*unfolding*”) a térképfelvételt végző program automatikusan elvégzi. A további vizsgálatokban ezzel a kétdimenziós, a felvett pontok számával megegyező sort és a hullámszámok számával egyező oszlopot tartalmazó $((n \times m) \times \lambda)$ mátrixszal dolgozhatunk.

Alapvonal-korrekción:

A látható tartományba eső gerjesztő fény a mintákban kölcsönhat a rezgési energiaszintekkel, azonban emellett fluoreszcenciából származó emissziót is okozhat. Ezért a Raman-spektrumok jellemzője, hogy az aktív Raman-sávok egy anyagra jellemző nagyságú fluoreszcens háttérre szuperponálódnak. Ezt a háttérrel minden térképen alapvonal-korrekciónal távolítottam el. Erre egyik megoldás szerint a spektrumra szakaszosan lineáris függvényt illesztettem majd ezt kivontam a spektrumból. Az összes térkép összes spektrumához, valamint a referenciaspektrumokhoz is ugyanazt az alapvonalat használtam.

A módszer nem volt alkalmazható abban az esetben, ha a komponenseknek nagyon eltérő volt az alapvonala. Ennek a problémának megoldására az Eilers-féle, aszimmetrikus legkisebb négyzetek módszere szerinti alapvonal-illesztést [86] alkalmaztam. Az eljárás két paraméter beállítását igényli: a p paraméter az illesztett alapvonal aszimmetriájáért felel (értéke általában 0,001 és 0,1 között van), Azt lehet ezzel beállítani, hogy az alapvonal mennyire metszhesse a korrigálni kívánt spektrumot, azaz mennyire „haladjon alatta”. A λ paraméter a simítási paraméter, mely értékére $10^2 \leq \lambda \leq 10^9$ –t javasolnak a módszer kifejlesztői. Az optimális paramétereket többszöri szisztematikus beállítással, empirikusan állapítottam meg.

Normálás:

A térképek pontjai közt spektrumról spektrumra intenzitásingadozások figyelhetők meg, mely a fókuszálás hibájából valamint a térképezett mintafelület egyenetlenségeiből adódnak. Ezt a spektrum (görbe alatti területére történő) normálásával küszöbölhetjük ki, amivel az intenzitásértékeket azonos skálára helyezzük, a következő képlet szerint.

$$x_{ij}^{\text{norm}} = \frac{x_{ij}}{\sum_{j=\lambda_1}^{\lambda_2} x_{ij}} \cdot 100, \quad (3.4.)$$

ahol x_{ij} az i -edik térkép-spektrum intenzitása a j -edik hullámszámváltozónál, λ_1 és λ_2 pedig a normáláshoz választott hullámszámtartomány szélei (a változók sorszámaként megadva).

Amennyiben a komponensek Raman-aktivitása között jelentős eltérés van, akkor a spektrumok közötti intenzitáskülönbségek az eltérő komponens összetételről is fontos információt hordoznak, amely a normálással jelentősen torzul és gátolja a pontos mennyiségi információk kinyerését.

Átlagra centrálás:

Az átlagra centrálás az eredeti intenzitásskálán való konstans eltolást jelenti oly módon, hogy az egyes hullámszámhoz tartozó intenzitásértékek számtani közepe 0 legyen. Lényegében tehát az adathalmaz összes spektrumából kivonjuk az adathalmaz átlagspektrumát.

$$x_{ij}^{\text{centr.}} = x_{ij} - \bar{x}_j; \quad \bar{x}_j = \frac{1}{r} \sum_{i=1}^r x_{ij}, \quad (3.5.)$$

ahol x_{ij} az eredeti adatmátrix i -edik sorának és j -edik oszlopának eleme, $\bar{x}_j$ a j -edik oszlop intenzitásértékeinek átlaga, r pedig a spektrumok száma.

Standardizálás:

A standardizálás konstans eltolást és zsugorítást jelent oly módon, hogy az elemek számtani közepe 0, szórás pedig egységnyi legyen.

$$x_{ij}^{\text{std.}} = \frac{x_{ij} - \bar{x}_j}{s_j}; \quad s_j = \sqrt{\frac{1}{r-1} \sum_{i=1}^r (x_{ij} - \bar{x}_j)^2} \quad (3.6.)$$

ahol x_{ij} az eredeti adatmátrix i -edik sorának és j -edik oszlopának eleme, az s_i pedig az i -edik tulajdonságvektor elemeinek tapasztalati szórása. A standardizálás során a változók dimenziómentessé válnak.

Ezt a műveletet a sorvektorokkal végezve sorstandardizálásról (SNV, *standard normal variate*) beszélünk.

Léptékváltás kitűzött minimum és maximum közé:

A léptékváltás konstans eltolást és zsugorítást (vagy nyújtást) jelent az eredeti skálán, mely során a változók kitűzött alsó és felső határ közé helyezhetők.

$$x'_{ij} = A_j + (F_j - A_j) \frac{x_{ij} - \min(x_{ij})}{\max(x_{ij}) - \min(x_{ij})} \quad (3.7.)$$

ahol x_{ij} az eredeti adatmátrix i -edik sorának és j -edik oszlopának eleme, A_j a kívánt alsó, F_j a felső határ, $\min(x_{ij})$ és $\max(x_{ij})$ pedig a vektor legkisebb illetve legnagyobb eleme. Ily módon történő léptékváltásra akkor lehet szükség, ha az alkalmazni kívánt függvény értékkészlete eltér az adathalmaz értékkészletétől.

Multiplikatív szóródási korrekció (MSC):

Az egyes spektrumok intenzitásingadozásainak kiküszöbölésére alkalmazható a multiplikatív szóródási korrekció, mely során elvégezzük minden spektrum átlagspektrummal való lineáris regresszióját:

$$x_i = a_i + b_i \bar{x}_j + e_i \quad (3.8)$$

Ahol x_i az i . korrigálni kívánt spektrum, $\bar{x}_j$ az átlagspektrum, e_i pedig a regresszió hibatagja, a_i az egyenes tengelymetszete, b_i a meredeksége. A korrigált spektrum a következő módon számítható:

$$x_{i, MSC} = (x_i - a_i) / b_i \quad (3.9.)$$

3.6. Alkalmazott regressziós módszerek

3.6.1. Egyváltozós regresszió

A Raman-spektrumok intenzitásértékei elméletileg lineárisan függenek a koncentrációtól a Lambert-Beer törvény értelmében. Ez alapján lehetőség van egy megfelelően kiválasztott hullámszámnál leolvasott intenzitásértékek és a koncentrációk közt lineáris regressziót végezni a következők szerint:

$$y = \alpha + \beta(x - \bar{x}) + \varepsilon \quad (3.10.)$$

$$\bar{x} = \sum \frac{x_i}{n} \quad (3.11.)$$

ahol α és β a becsülendő paraméterek, α az egyenes tengelymetszete, β a meredeksége, ε pedig a hiba.

3.6.2. Klasszikus legkisebb négyzetek módszere (CLS)

A legegyszerűbb többváltozós regressziós módszer, a klasszikus legkisebb négyzetek módszere (CLS, *Classical Least Squares*) az egyes spektrumokat a referenciaspektrumok lineáris kombinációiként becsüli a 3.12. egyenlet szerint. Eszerint a becsült spektrumok a megfelelő súlyértékekkel szorzott egyedi komponensspektrumok összegzéséből jönnek létre. A becsült súlyfaktorok a koncentrációknak feleltethetők meg, mely az egyenlet átrendezéséből kapható meg (3.13. egyenlet).

$$\mathbf{X} = \mathbf{CS}^T + \mathbf{E} \quad (3.12.)$$

$$\mathbf{C} = \mathbf{XS}(\mathbf{S}^T\mathbf{S})^{-1}, \quad (3.13.)$$

ahol $\mathbf{X}$ az adatmátrix ($n \times k$), $\mathbf{S}^T$ a komponensek referenciaspektrumaiból képzett mátrix ($p \times k$), $\mathbf{C}$ az egyes pontok spektrális koncentrációit tartalmazza ($n \times p$), $\mathbf{E}$ a hibamátrix ($n \times k$), mely négyzetes hibaértéket minimalizálja az algoritmus).

A módszer előnye, hogy nagyon egyszerű és a referenciaspektrumok ismeretében kalibrációs sorozat nélkül is becsülhető a koncentrációprofil. Ugyanakkor ezáltal a módszer lényegében kétpontos kalibráció, ami legtöbbször eléggé pontatlan becsléshez vezet, ezért a nyert eredmények maximum félkvantitatív információként kezelhetők. Kalibrációs sor alkalmazásával lehetőség nyílik arra, hogy az algoritmus maga alkossa meg a modellezéshez használt referenciaspektrumokat a kalibrációs sor spektrumainak segítségével (3.14. egyenlet szerint), majd ezeket használja fel a súlyfaktorok számításához.

$$\mathbf{S}^T = (\mathbf{C}^T\mathbf{C})^{-1} \mathbf{C}^T\mathbf{X} \quad (3.14.)$$

Ez olyan esetekben előnyös lehet, mikor nem áll rendelkezésre a tiszta komponensek referenciaspektruma, vagy feltételezhető, hogy a keverékben fellépő fizikai vagy kémiai hatások következtében az egyes komponensek spektruma eltér a tiszta mintában mérhetőtől.

3.6.3. Főkomponens-analízisen alapuló regressziós módszerek- Főkomponens-regresszió (PCR) és részleges legkisebb négyzetek módszere (PLS) [25]

A főkomponens-regresszió (PCR, *Principal Component Regression*) és a részleges legkisebb négyzetek módszere egyaránt főkomponens-analízisen alapuló regressziós módszer. Főkomponens-analízis során az eredeti változótérből (jelen esetben ezer hullámszám-csatornából felépített ezerdimenziós tér) matematikai transzformáció (szinguláris-érték felbontás)

segítségével olyan új változókat hozunk létre, melyek a legjobban leírják az adathalmazban lévő varianciát. Ezek a változók a komponensspektrumokból kombinálódnak, ortogonálisak és az új, szűkített dimenziós térünk bázisai. A modellekhez felhasznált látens változók/főkomponensek számát mindig a felhasználó dönti el, mert csak az első néhány új változóban (főkomponensek) lesz az információ nagy része, amely a regresszióhoz szükséges. Ezzel a módszerrel az adathalmaz dimenziója jelentősen csökkenthető. A regressziót ezután a redukált adathalmazzal lehet elvégezni. Mindegyik főkomponenshez adott főkomponens-együtthatók rendelkeznek, amelyek szorzatösszegeként a spektrumokat kapjuk meg.

A részleges legkisebb négyzetek módszere annyiban különbözik a főkomponens-regressziótól, hogy az új változók (látens változók) úgy jönnek létre, hogy a függő változóval való kovarianciát maximalizálják. Ezáltal a látens változók jobban korrelálnak a függő változóval, így a meghatározási hiba csökkenése várható a módszer alkalmazásával.

3.6.4. Változókiválasztási módszerek

A Raman-spektrumok hullámszámaint felfoghatjuk változóként, ahogy ezt a kemometriai módszerek is teszik. Egy spektrum több száz változót tartalmazhat, de nem mindegyik hullámszám hordoz lényegi információt a mintában lévő komponensekről. A kevésbé informatív hullámszámoknál mért intenzitások sokszor a kiértékelés szempontjából feleslegesek, sőt ronthatják a mennyiségi meghatározás pontosságát. Bár a főkomponens-analízisen alapuló módszerek esetén is dimenziócsökkentést hajtunk végre, a főkomponensek/látens változók megalkotása az összes változó figyelembevételével történik, ennek következményeként a nem informatív változók is kifejtik hatásukat. Ezért célszerű, ha ezeket a változókat modellépítés előtt eltávolítjuk az adathalmazból. A spektrumok esetén azt is figyelembe kell venni, hogy a szomszédos változók nem függetlenek egymástól, mivel az egyes csúcsok/sávok szélesebb hullámhossztartományra is kiterjednek. Így ahhoz, hogy gyakorlati szempontból is értelmezhető legyen a változókiválasztási módszer, tartományok kiválasztására kell törekedni. Munkám során két változókiválasztási módszert, az intervallum- PLS-t és a genetikus algoritmusokat teszteltem.

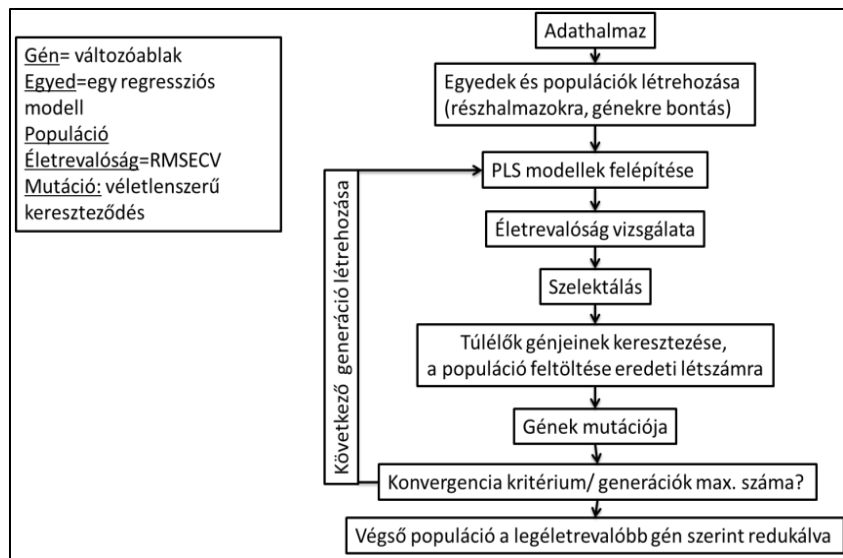
3.6.4.1. Intervallum-PLS [66]

Az intervallum-PLS a spektrumokat változóablakokra bontja fel úgy, hogy az egyes ablakok általunk definiált számú, egymással szomszédos változókat tartalmaznak. Az algoritmus kétféleképpen futtatható. Beállítástól függően lehetséges lépésenkénti változóbevonás, amikor

egy kezdeti változóintervallumhoz adja hozzá a következőket (*forward iPLS*), amíg fel nem épül egy optimális számú változót tartalmazó modell. Fordított esetben változóelhagyásról beszélünk, mely esetén az összes ablakot tartalmazó kezdeti modellből folyamatosan hagyja ki azokat (*reverse iPLS*). Az eljárás során minden új ablak beépítése/ kivétele után az aktuális adathalmazból PLS modell készül és a modellek keresztellenőrzési hibájának (RMSECV) minimalizálásával dönti el, hogy mely változók modellbe való beépítése/benne hagyása javítja vagy rontja a meghatározást. Ilyen módon az iPLS szisztematikus keresés során határozza meg, hogy melyek az adathalmaz legfontosabb változói. A megfelelő modell megtalálása optimálást igénylő feladat.

3.6.4.2. Genetikus algoritmusokkal végzett változó kiválasztás [63]

Az intervallum-PLS hátránya, hogy a szisztematikus keresés következtében gyakran csak a lokális minimumot találja meg. A módszernek ezt a hátrányát globális minimumkeresési módszerrel, például genetikus algoritmusok alkalmazásával küszöbölhetjük ki. A módszer a természetből vett evolúció analógiáját használja fel arra, hogy globális minimum értéket találjon. A változóablakokat (*gének*) véletlenszerűen választja ki a változóablakok ki-bekapcsolásával (0 és 1 értékek rendelése a változóablakokhoz) majd ezekből alakítja ki a regressziós modellt („*egyed*”), melyek alkotják a *populációkat*. A felépített modellek RMSECV értékeit („*életrevalóság*”) vizsgálva dönt az egyes halmazok elhagyásáról illetve megtartásáról („*szelekció*”), a keresés pedig a megadott konvergenciakritérium vagy a maximális generációs szám elérésével áll le. Az algoritmus futásának menetét a 3.6.1. ábra szemlélteti. A genetikus algoritmusok hátránya, hogy a véletlen keresés miatt a kiválasztás eredménye nem reprodukálható.



3.6.1. ábra: A genetikus algoritmus futásának vázlata

3.6.5. Nemlineáris módszerek

3.6.5.1. Polinomiális PLS

A PLS regresszió a nemlinearitás kezelésére úgy tehető képessé, ha a főkomponens-együtthatókkal végzett regresszió során magasabb fokú regressziós függvényt alkalmazunk a lineáris helyett. A polinomiális PLS modellhez ilyenkor a látens változók számának meghatározása mellett az illesztett polinom fokszámát is optimalni kell.

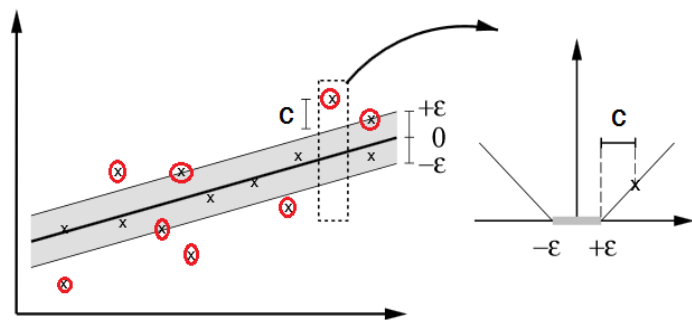
3.6.5.2. Helyileg súlyozott regresszió (LWR) [71]

Helyileg súlyozott regresszió (LWR, *Locally Weighted Regression*) során úgy illesztünk nemlineáris görbét a kalibráló pontokra (melyek lehetnek PLS-el nyert főkomponens-együtthatók is), hogy a regresszió során nem az összes kalibrációs pontot vesszük figyelembe, hanem az adott pontnak csak bizonyos számú szomszédját. Az LWR a szomszédos pontokat az éppen meghatározandó ponttól vett távolsággal súlyozza is, ezért a távolabbi pontok kisebb súllyal vesznek részt az adott pont körüli illesztésbe. A súlyozás történhet például trikörös súlyfüggvénnyel. Az alkalmazás szempontjából legfontosabb kérdés a szomszédosnak tekintett pontok optimális számának megállapítása. A módszer fő előnye, hogy rugalmasan kezeli a nemlinearitást, nem kell előre megadni az illesztett polinom fokszámát.

3.6.5.3. Támogatóvektorok módszerével végzett regresszió (SVM) [81,82]

A támogatóvektorok módszere olyan gépi tanulási algoritmus, mely az adatmátrixot magasabb dimenzióba transzformálja úgy, hogy ezáltal a nemlineáris adatok lineárisra illeszkedjenek. A transzformációt egy ún. „kernelfüggvény” segítségével valósítja meg, mely „kacsaringósságát” a γ paraméter jelzi.

A regressziós függvény illesztéséhez egy ε paraméterrel definiált nagyságú tűréshatárt használunk, melyen belül a pontok a függvény illesztésében nem vesznek részt, csak a tűréshatártól távolabb eső pontokat „bünteti”, a



3.6.2. ábra: Támogatóvektorok módszerével végzett regresszió. A piros körrel jelzett pontok a támogatóvektorok [87]

távolsággal egyenes arányban. Az arányossági paraméter a C (cost) veszteségi paraméter. A kívülre, illetve a tűréshatárra eső pontok a támogatóvektorok, melyek kifeszítik a regressziós függvényt (3.6.2. ábra).

3.6.5.4. Mesterséges ideghálózatok

A mesterséges ideghálózatok (ANN, *artificial neural networks*) az emberi agy tanulási folyamatát, információfeldolgozását utánozva kezelnek nemlineáris sokváltozós folyamatokat. A modellezés során a bemeneti egységeket (*input*) egy vagy több feldolgozó, rejtett csomópont (*hidden node*) alakítja át kimenetűvé (*output*). A rejtett réteg(ek) (*layer*) tevékenységét a rétegek kapcsolataihoz rendelt súlytényezők (w) illetve a rejtett réteg átalakító függvénye befolyásolja, mely általában lineáris, tangens hiperbolikus vagy szigmoid függvény.

Regresszióra az ún. „felügyelt” tanítású mesterséges ideghálózatok alkalmasak. Ebben az esetben a bemeneti és kimeneti párok ismertek és a hálózat betanítására használják, azaz a bemenetre kapott válasz összehasonlítható a kívánt kimenettel. A bemeneti egység a kalibráló minták spektrumait tartalmazza, míg a kimeneti egységek a komponensek koncentrációértékei. A kalibráló halmazt az algoritmus tanító, validáló és tesztelő halmazokra bontja fel, a validáló halmazt használva a megfelelő hálózat felépítéséhez. Ennek a felbontásnak az aránya beállítástól függ. Az eljárás paramétereit ehhez az értékekhez lehet illeszteni, és a kedvező irányba módosítani. A betanításra számos algoritmus ismert, munkám során ezek közül a visszafuttatásos (*back-propagation*) algoritmust alkalmaztam. Ez a kimenő adatokat a kívánt értékkel összehasonlítja, a hibaértékek pedig a hálózaton visszafelé terjedve módosítják a súlytényezőket (melyek véletlen számként indultak) úgy, hogy a hiba csökkenjen. A rejtett rétegben lévő közbülső csomópontok száma jelentősen befolyásolja a modellezés sikerességét. Túl kevés csomópont alkalmazása esetén a hálózat nem jellemzi megfelelően a rendszert, túl sok esetén azonban túlillesztés lép fel, melyre a módszer különösen hajlamos.

A mesterséges ideghálózatok futtatására számos algoritmus és regularizációs/szabályozó paraméter áll rendelkezésre, melyek bemutatása e dolgozatnak nem célja. Spektroszkópiai szemszögből fontos aspektusairól részletesebb leírás található például Pérez-Marin munkájában [72].

3.7. Homogenitás-vizsgálat a makropixel-elemzés módszerével [49,50]

Raman-térképezés során felmerül a kérdés, hogy mekkora az a térképméret, mely már megfelelően reprezentatív ahhoz, hogy a minta átlagos koncentrációjáról tudjunk következtetéseket levonni. Ennek megválaszolására adhat lehetőséget a makropixel-elemzés módszere.

Makropixel-elemzés során az eredeti térképet egy randomizált térképpel hasonlítjuk össze, melyet a térkép pontjainak véletlenszerű átrendezésével (randomizálással) állítunk elő. Ezután mindkét térképet felosztjuk különböző nagyságú makropixelexekre („kisebb térképekre”). Az egyes makropixelexek szórásaiból számítható a makropixelexek szórásértékeinek számtani közepe külön-külön a randomizált és valódi térképre, a 3.15.-3.17. egyenletek szerint. Ezek hányadosa a homogenitást jellemző Poole-indexet adja (3.18. egyenlet).

$$\bar{X}_i = \frac{1}{n} \sum_{j=1}^n x_{ij} \quad (3.15.)$$

$$S_i = \frac{1}{\sqrt{n-1}} \sqrt{\sum_{j=1}^n (x_{ij} - \bar{X}_i)^2} \quad (3.16.)$$

$$\bar{M}_{S_i} = \frac{\sum_{i=1}^m S_i}{m} \quad (3.17.)$$

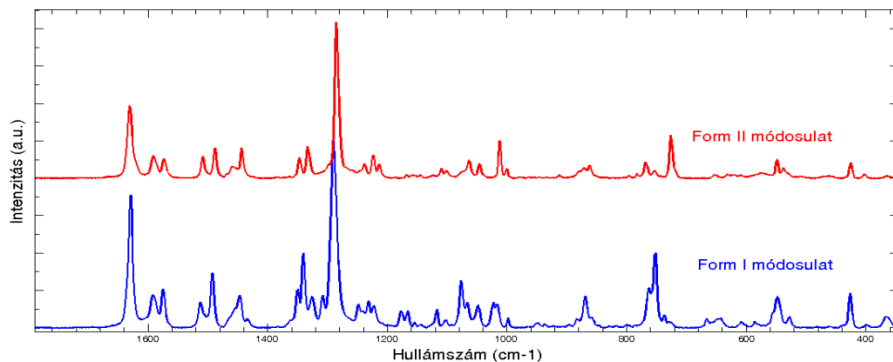
$$M_P = \frac{\bar{M}_{S_{i, \text{valódi}}}}{\bar{M}_{S_{i, \text{random}}}} \quad (3.18.)$$

,ahol $\bar{X}_i$ az i -edik makropixelben lévő értékek számtani közepe, n az egy makropixelben lévő pontok száma, x_{ij} az i -edik makropixelben lévő j -edik pont értéke, az i -edik makropixel szórása, S_i a makropixelexek szórásértékeinek számtani közepe, m a makropixelexek száma, $\bar{M}_{S_{i, \text{valódi}}}$ a valódi térkép makropixel szórásértékeinek számtani közepe, $\bar{M}_{S_{i, \text{random}}}$ a randomizált térkép makropixel szórásértékeinek számtani közepe, a számított M_P Poole index. A Poole-index 0 és 1 közötti értékeket vesz fel, az 1 érték jelenti azt, hogy a vizsgált térképen a koncentrációk teljesen random eloszlásúak.

4. Eredmények

4.1. Polimorf rendszer mennyiségi elemzése

A vizsgálataimhoz carvedilol polimorfok keverékrendszerét alkalmaztam, amelyen egyszerűségének köszönhetően jól össze lehet hasonlítani a különböző módszerek előrejelző képességét, ugyanakkor a polimorf rendszerek esetén felmerülő stabilitási problémák is tanulmányozhatók segítségével. A két polimorf módosulat spektrumában nagyon kevés eltérés tapasztalható (4.1.1. ábra). A hagyományos módszerekkel való értékelésük problémákba ütközhet, ezért is kiemelten fontos ebben az esetben hatékony többváltozós adatelemzési módszereket alkalmaznunk, amelyekkel a legtöbb információt össze lehet gyűjteni a komponensek Raman-spektrumaiból.



4.1.1. ábra: A két carvedilol polimorf módosulat Raman-spektruma

A kvantitatív modellek építéséhez a 19 keveréket tartalmazó mintasorozatot 12:7 arányban felosztottam kalibráló és validáló halmazra. A validáló halmazba kerültek a 2%, 5%, 11%, 50%, 89%, 95%, 98% Form I-et tartalmazó keverékek, míg a többi a kalibráló halmazt alkotta (ld. 3.1.1. táblázat). Három elem kihagyásos keresztellenőrzést végeztem, azaz egyszerre egy egész koncentrációsintet hagytam ki (3 ismételt mérést).

4.1.1. Adatelőkészítési módszerek vizsgálata

A Raman spektrumok jellegzetessége a fluoreszcens sugárzásból eredő háttér jelenléte, mely nem hordoz kémiai információt. Ezért a megfelelő kiértékelés érdekében szükséges ezt eltávolítani, melyet alapvonal-korrekcióval végeztem. Az összes térkép összes spektrumára ugyanazt a szakaszosan lineáris függvényt illesztettem, mivel így az alapvonalbeli eltérések nem befolyásolhatják a modellezést. Szintén szükséges volt normálást végezni, a térképezés során

felmerülő intenzitásingadozások kiszűrésére. Ez a polimorfok esetén nem okoz jelentős információvesztést a komponensek nagyon hasonló, ugyanolyan intenzitású spektrumai miatt.

Ezen szükséges adatelőkészítéseken kívül fontos feladatnak tekintetem megvizsgálni egyéb módszerek modelljóságra gyakorolt hatását is minden regressziós technika esetén. Tanulmányoztam az átlagra centrálás, a standardizálás, a sorstandardizálás (SNV) és a multiplikatív szóródási korrekció (MSC) alkalmazhatóságát. Egy ilyen vizsgálatot a PLS modell esetén mutatok be részletesen, mely talán a legszélesebb körben alkalmazott kemometriai módszer. A PLS modellek építését részletesebben a 4.1.2.3. pontban tárgyalom. Az összes többi módszer esetén hasonlóképpen határoztam meg, hogy milyen adatelőkészítéssel nyerhető a legpontosabb modell és a továbbiakban csak ezeket mutatom be.

A különféle adatelőkészítések PLS modellre gyakorolt hatását a 4.1.1. táblázatban foglalom össze. A meghatározás közepes négyzetes hibáinak összehasonlítása alapján megállapítható, hogy a Raman-térképeken rutinszerűen használt alapvonal-korrekció és normálás mellett más adatelőkészítési módszerek használatával jelentősen csökkenthető a meghatározás hibája. A PLS módszer esetén legjobbnak a centrálást és a sorstandardizálást egyaránt használó modell bizonyult, mellyel az RMSE értékek közel a felére csökkentek a csak alapvonal-korrekciót és normálást használó modellhez képest.

4.1.1. táblázat: Az alkalmazott adatelőkészítések hatása 2 látens változót használó PLS modellekre. (alapv.=alapvonal-korrekció, MSC= multiplikatív szóródási korrekció, SNV=sorstandardizálás)

	alapv., normálás	alapv., normálás, centrálás	alapv., normálás, standardizálás	alapv., normálás, centrálás, MSC	alapv., normálás, centrálás, SNV
RMSEC	3,69	1,82	1,81	2,38	1,68
RMSECV	4,16	2,81	2,68	3,36	2,16
RMSEP	2,94	2,27	2,24	2,98	1,91

4.1.2. Referenciamódszerekkel felállított modellek

4.1.2.1. Egyváltozós regresszió

Az egyváltozós regresszió használatának feltétele, hogy a spektrumban az adott komponensre szelektív csúcsot, vagy sávot tudjunk találni, amelynek intenzitása vagy a görbe alatti területe könnyen meghatározható és a koncentrációval jó linearitást mutat. Polimorfok

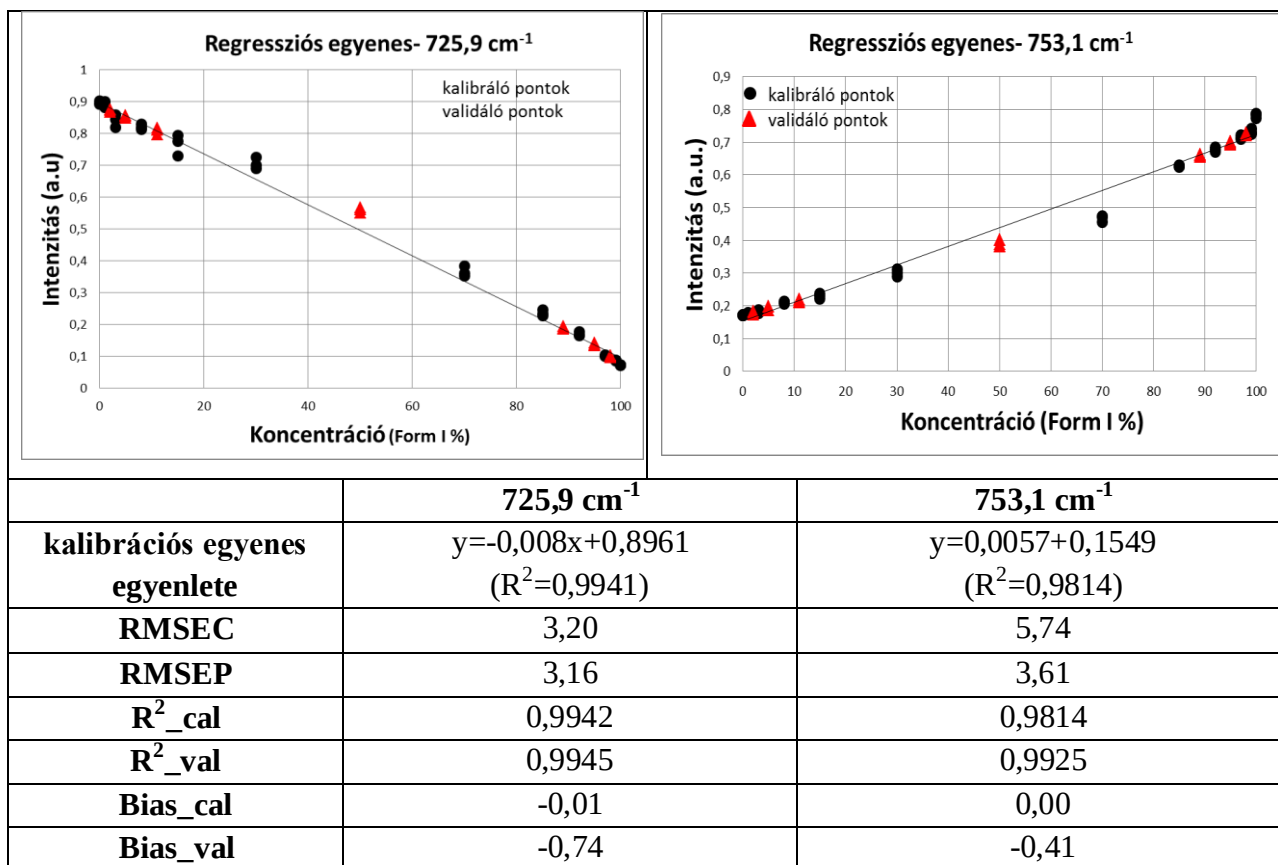
rendszerében a szelektivitás az erősen átfedő csúcsok miatt gyakran nem teljesül, valamint a lineáris viselkedést is módosíthatják az anyagban fellépő kölcsönhatások és a 3. fejezetben említett adatelőkészítési módszerek (elsősorban a normálás).

A vizsgált modellrendszerünkben két viszonylag szelektív csúcs található a $725,9\text{ cm}^{-1}$ –nél (Form II polimorf módosulatra jellemző csúcs) és a $753,1\text{ cm}^{-1}$ -nél (Form I-re jellemző csúcs). Mindkét hullámszámnál leolvasott intenzitásértékekkel elvégeztem a lineáris regressziót.

A kapott eredmények alapján (4.1.2. táblázat) megállapíthatjuk, hogy a Form II módosulatra jellemző $725,9\text{ cm}^{-1}$ -es csúccsal végzett regresszió pontosabb eredményeket szolgáltat. A két intenzitásértéken végzett regresszióval nyert eredmények jelentősen különböznek egymástól. Ennek oka az lehet, hogy a $753,1\text{ cm}^{-1}$ -nél lévő csúcs kevésbé szelektív, a Form II-nek is megjelenik ott csúcsa. A módszer további hátránya, hogy az egyváltozós regressziók sokszor nem elég robusztusak: a mennyiségi meghatározást jelentősen befolyásolja, hogy melyik hullámszámnál végezzük a kiértékelést és zavarhatják kiugró adatok, véletlen hibák (zajok) vagy az alapvonal-illesztésből adódó pontatlanságok.

Bár a becsült koncentrációértékekből számolt négyzetes hiba a 3%-ot közelítette meg, mely „elég jónak” mondható, bizonyos esetekben szükség lehet a pontosabb meghatározásra. A polimorf keverékek kvantitatív vizsgálatára például legtöbbször kis mennyiségű (1-5%) szennyező meghatározásához van szükség. Ha a Raman-kémiai térképezéses technikát kis koncentrációk pontos kimutatására szeretnénk használni mint a PAT és a minőségbiztosítás egyik eszköze, akkor a modelleknek sokkal pontosabb becslést kell adniuk a szennyezők koncentrációjára. A felsorolt tapasztalatok alapján tehát kijelenthetjük, hogy többváltozós regressziós módszerek alkalmazására feltétlenül szükség van ahhoz, hogy a kvantitatív meghatározásaink megfelelően pontossá váljanak.

4.1.2. táblázat: Egyváltozós lineáris regresszió eredményei



4.1.2.2. Klasszikus legkisebb négyzetek módszerével végzett többváltozós regresszió

A legegyszerűbb többváltozós regressziós módszer, amit alkalmaztam a klasszikus legkisebb négyzetek módszere volt. Ennek során lehetőségem volt a tiszta komponensek spektrumainak segítségével becsülni az egyes minták koncentrációit. Ezt összehasonlítottam a kalibráló sorból számolt referenciaspektrumokat használó regresszióval.

A carvedilol polimorfok fizikai keverékeinek kvantitatív meghatározása során nem tapasztaltam jelentős különbséget az első („CLS modellezés”) és a második („CLS regresszió”) lehetőség közt, illetve nem dönthető el egyértelműen, hogy melyik módszer szolgáltat pontosabb eredményt. (4.1.3. táblázat)

A módszer nem volt pontosabb, mint az egyváltozós lineáris regresszió, egyes koncentrációértékek értékelésekor pedig nagy félrebecsléseket adott. Ennek oka a rendszerben jelenlévő nemlinearitás lehet, melyet a reziduum ábra is jól szemléltet (4.1.3. ábra).

Az eredményeink azt támasztják alá, hogy a klasszikus legkisebb négyzetek módszere alapján végzett kiértékelés a polimorf keverékek gyors, de csupán közelítő becslésére illetve a minták összehasonlítására használható.

4.1.3. táblázat: CLS módszer eredményei

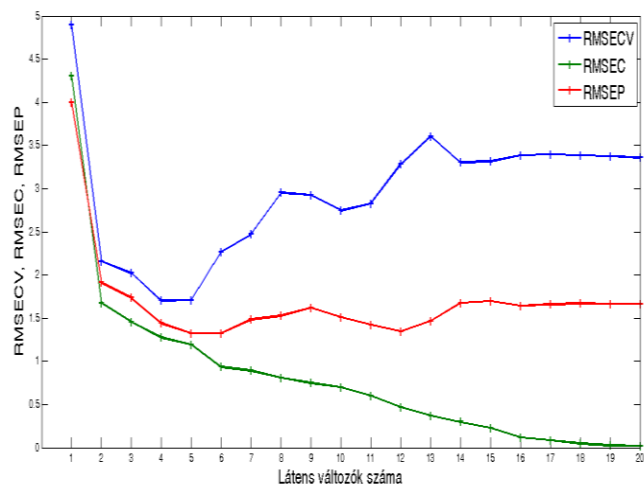
	CLS modellezés	CLS regresszió
adatelőkészítés	alapvonal-korrekció, normálás	
RMSEC	3,48	3,70
RMSECV	-	4,19
RMSEP	3,63	3,00
R²_cal	0,9947	0,9922
R²_CV	-	0,9900
R²_val	0,9952	0,9950
Bias_cal	-0,02	0
Bias_CV	-	0,14
Bias_val	-0,03	-0,42

4.1.2.3. Főkomponens-regresszió és részleges legkisebb négyzetek módszere

Ahogy azt a 3.6.3. pontban említettem a regressziós módszerek általános bemutatása során, mindkét módszer alkalmazása során a kulcskérdés az alkalmazott főkomponensek/látens változók számának meghatározása. Azt kell megállapítanunk, hogy hány főkomponens tartalmazza a számunkra releváns információt, illetve melyek azok a főkomponensek, amelyekben már főleg csak zavaró hatások érvényesülnek.

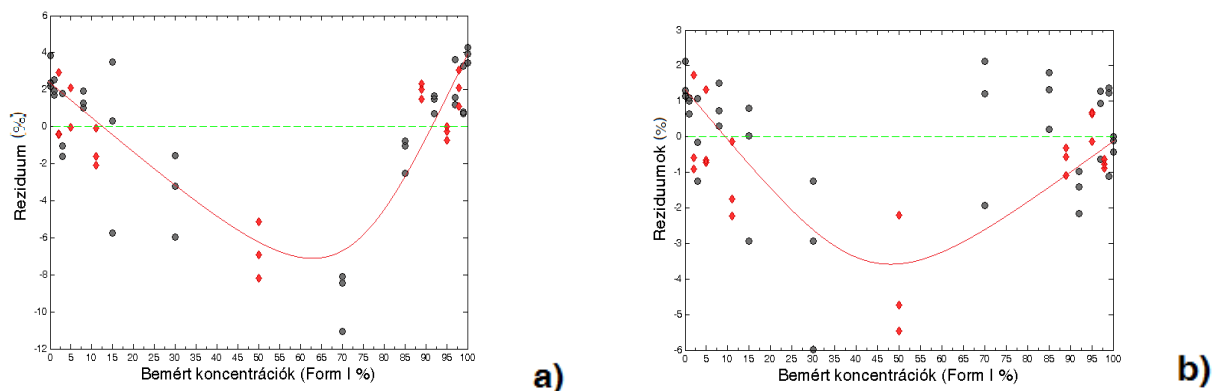
Mivel a fő célunk minél jobb előrejelző képességű modellek építése, ezért a főkomponensek számának meghatározása során a keresztellenőrzési hiba minimalizálására törekedtem. Megvizsgáltam a különböző számú főkomponenseket használó modelleket, és azt a modellt választottam megfelelőnek, melynél az RMSECV és RMSEP értékek minimálisak (4.1.2. ábra). Ugyanakkor az is ismert, hogy a vizsgált keverékrendszer két összetevőt tartalmaz. Ha kettőnél sokkal több főkomponenst választunk, akkor szinte biztos, hogy a véletlen zajra is illeszttem a modellt, amellyel a prediktív erő romlik. Ezeket a szempontokat mérlegeltem és a két főkomponenssel/látens változóval végzett PCR és PLS regressziót tekintettem optimálisnak.

A nyert eredményeket a CLS modellezéssel összehasonlítottam (4.1.4. táblázat). Megállapítottam, hogy jelentős javulás érhető el az adatmátrix dimenziójának csökkentésével. Megfelelő adatelőkészítést alkalmazva a CLS 3-4%-os meghatározási hibájával szemben PLS regresszióval már 2% alá sikerült csökkenteni mind a kalibráció, mind a predikció hibáját. Azt az elméleti megfontolást is sikerült alátámasztani, hogy a részleges legkisebb négyzetek módszere kisebb hibával dolgozik, mint a főkomponens-regresszió.



4.1.2. ábra: PLS regresszió látens változó számának meghatározása RMSE értékek alapján.

Az eddig vizsgált módszerek hátránya, hogy nem kezelik a gyakorlatban gyakran megjelenő nemlinearitást. Nemlineáris viselkedés felléphet mintabeli illetve készülékbeli sajátságok miatt valamint egyes adatelőkészítési módszerek (pl. normálás) is belevihetnek a rendszerbe. Az általam vizsgált modellrendszerben is fellépett nemlinearitás, melyet a lineáris modellek reziduum-ábráin piros vonallal jelöltem (4.1.3. ábra). A nemlineáris viselkedés kezelése feltétlen fontos a megfelelő mennyiségi meghatározás érdekében. A továbbiakban olyan modellek felállítására törekedtem, melyek csökkentik, illetve kiküszöbölik ezt a jelenséget.



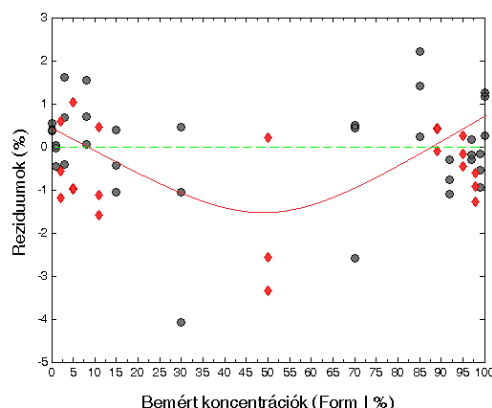
4.1.3. ábra: a) CLS modell reziduumai b) PLS modell reziduumai.

4.1.3. Változókiválasztási módszerekkel támogatott mennyiségi vizsgálatok

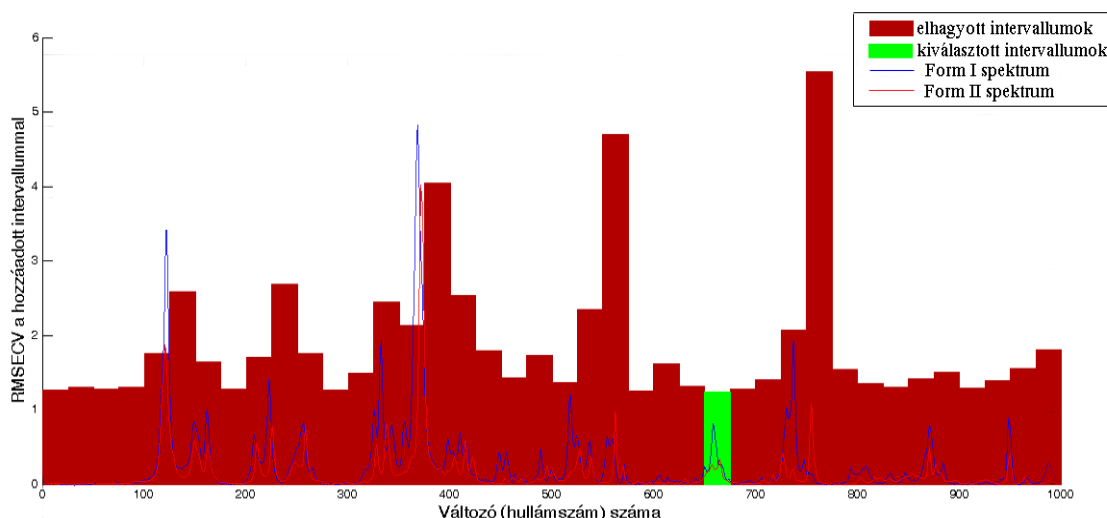
Hatóanyag-polimorfok esetén általában nagyon hasonló spektrumokat kell kezeljünk. A különbségek kiemelésével a regressziós modell pontosabbá válását várhatjuk, éppen ezért fontos lehet a változókiválasztások alkalmazása. A carvedilol polimorf rendszer mennyiségi meghatározása során tanulmányoztam ennek lehetőségét az intervallum-PLS és a genetikus algoritmusok használata során. Megvizsgáltam a különböző nagyságú változóablakok használatának hatását valamint az intervallum-PLS esetén a két futási formát is összevetettem.

Megállapítottam, hogy mind a két módszer eltávolította a változók jelentős részét, azaz helyes az a feltételezés, hogy a polimorfok hasonló spektruma miatt nem érdemes az egész spektrumot használni a mennyiségi meghatározás során. Az intervallum-PLS esetén megfigyeltem, hogy a két futási módszer nem ugyanazokat a változókat választja ki, hanem különböző lokális RMSE minimumokat talál. Ugyanazon változóablak nagyság alkalmazása esetén a *reverse* mód esetén többször figyeltem meg túlillesztést, melyre a változókiválasztások alkalmazása során különös tekintettel kell lenni. A változóablakok nagyságának vizsgálata során a modellezést 100, 50, 25, 10 és 5 változót tartalmazó ablakokkal végeztem el. Az eredményeim azt mutatták, hogy a meghatározás pontossága az ablak méretének csökkenésével is javítható, azonban ezt nem érdemes túlzott mértékben csökkenteni. A 10 és 5 változót tartalmazó ablakokat használó modelleknél már túlillesztettek lettek a modellek.

A 4.1.5. ábrán 25-ös ablakméretet használó forward iPLS modell futását követhetjük nyomon. Megfigyelhető, hogy az egyes komponensekre legszelektívebb csúcsok modellbe illesztése kiugróan magas keresztellenőrzési hibát mutatott, ezért ezek bevonása a modellekbe nem történt meg. A 25-ös ablakméretet használó forward iPLS modell eredményezte a legkisebb meghatározási hibát illetve a reziduum-ábrát (4.1.4. ábra) tekintve a nemlinearitás is nagymértékben csökkent. Ezek alapján arra következtethetünk, hogy esetünkben legjobban a szelektív csúcsok felelősek a nemlinearitásért, így megfelelő változókiválasztással a nemlinearitás is csökkenthető.



4.1.4. ábra: Az iPLS modell reziduumai

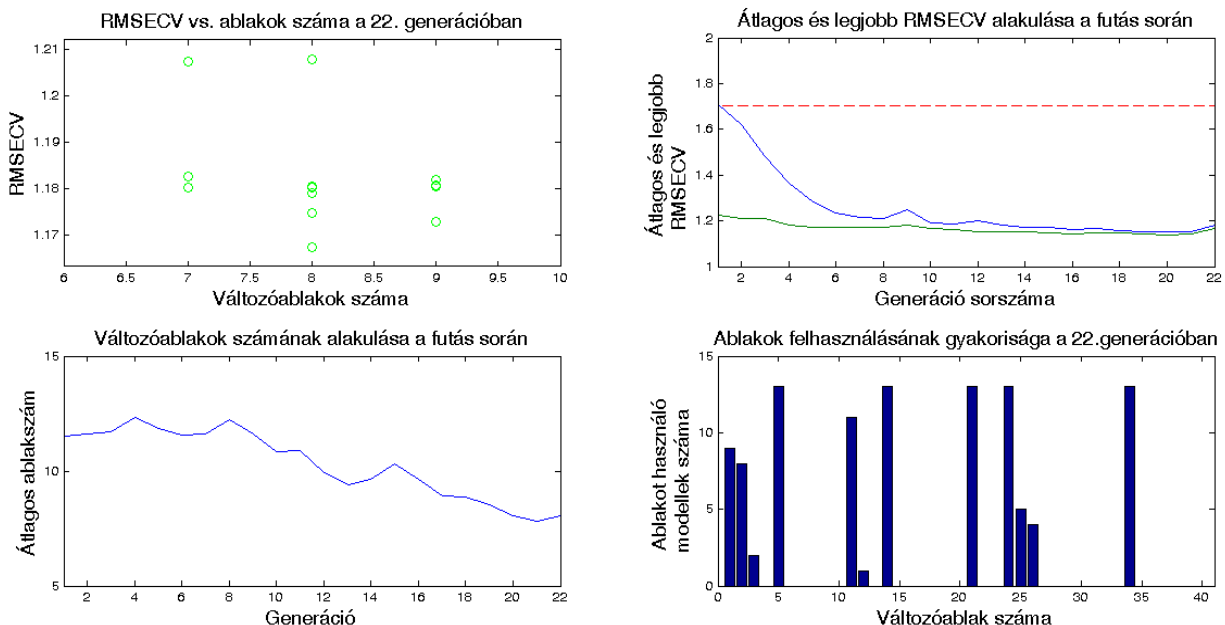


4.1.5. ábra: *Forward* iPLS: RMSECV értékek az egyes intervallumok modellbe való bevonásának során

Míg az intervallum- PLS alkalmazásával gyakran csak lokális minimumok megtalálására van lehetőség, a genetikus algoritmusok globális minimumkereső módszer. Hátránya azonban, hogy a véletlen keresés miatt a kiválasztás eredménye nem reprodukálható. Az algoritmus többszöri futtatásával azonban megfigyelhető volt, hogy a párhuzamos futtatás eredményeként kapott modellek hasonló eredményt szolgáltatnak és egyes változók ismételtelen megjelennek a modellekben. A populáció méretének és az ablakszélességnek meghatározása optimalizálási feladat. Munkám során a populációméretet 64-nek tartva vizsgáltam az 50, 25, 10 és 5 változót tartalmazó ablakméretekkel végzett modellépítések eredményét. Az algoritmus futását diagramokkal követhetjük nyomon (4.1.6. ábra). Ezekről leolvasható, hogy a generációk számának növelésével az életrevalóság nő, (azaz az RMSECV értékek csökkennek), illetve az algoritmus előrehaladásával egyre kevesebb ablakot tartalmaznak a modellek. A kisebb változóablakoknál ebben az esetben is túlillesztés figyeltem meg, így mindkét változókiválasztási módszer esetén a 25-ös ablakméretet tekintettem optimálisnak.

A két változókiválasztási módszert összehasonlítva azt állapíthatjuk meg, hogy mindkét módszerrel hasonló pontosságú modell építhető. Mind az iPLS mind a PLS-GA alkalmazásával a meghatározás az 1%-os hibahatárt közelítette meg (4.1.4. táblázat). Az RMSE értékek elemzése mellett pedig azt is figyelembe kell venni, hogy jelentősen csökkent a nemlineáris viselkedés. A redukált változós számú adathalmazokkal olyan modellt sikerült építeni, mellyel az egész koncentrációtartományban javult a meghatározás pontossága. Ezeken kívül a módszerek

alkalmazása különösen előnyös lehet kontrollált technológiák esetén. Ugyanis ha a kalibráció során kiderül, hogy mely változók szükségesek a pontos mennyiségi meghatározáshoz, ezek ismeretében tudatosabban lehet meghatározni a felvenni kívánt spektrális tartományt, így akár a mérési idő is csökkenthető. Ez folyamatos technológiák esetén kulcsfontosságú tényező lehet.



4.1.6. ábra: Genetikus futtatásának nyomon követési diagramjai

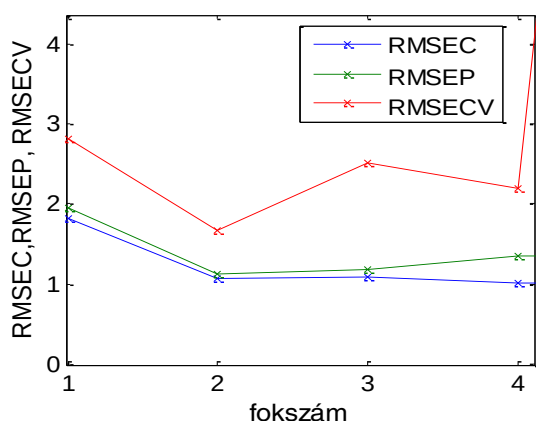
4.1.4. Nemlineáris regressziós módszerek tesztelése

Ahogy az előző fejezetben bemutattam, a változókiválasztási módszerek is csökkentették a rendszerben lévő nemlinearitást azzal, hogy a kevésbé nemlineáris viselkedést mutató változókat válogattuk ki. Azonban számos olyan regressziós módszer is létezik, melyek ezt a jelenséget figyelembe veszik és kezelik, ezáltal lehetőség van nemlineáris függvény illesztésére is. Dolgozatom következő szakaszában az ezekkel kapcsolatos vizsgálatokat és eredményeket mutatom be.

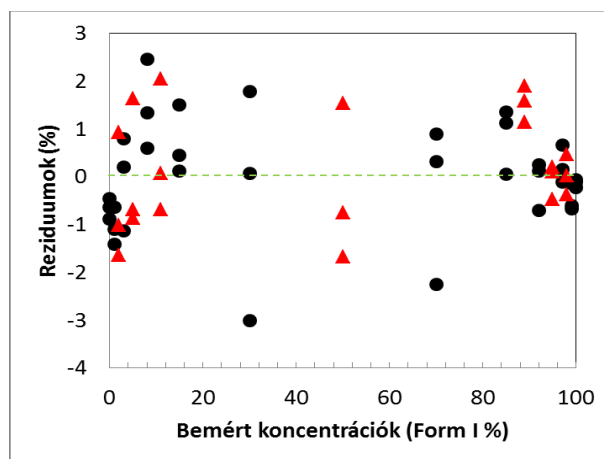
4.1.4.1. *Polinomiális részleges legkisebb négyzetek módszere*

Ha nemlinearitást tapasztalunk, akkor lehetőségünk van arra, hogy a PLS modellezés során magasabb fokú regressziós függvényt alkalmazzunk a lineáris helyett. Ezt a módszert alkalmazva tehát nem csupán a látens változók számát kell meghatározni, hanem a regressziós polinom fokszámát is.

A látens változók számát a lineáris PLS modellnél meghatározott kettőnek választottam, a polinom fokszámának kiválasztása pedig az RMSE értékek vizsgálata alapján történt. A különböző fokszámú polinomokkal végzett PLS modellezések RMSE értékei a polinom fokszámának függvényében jelentősen megváltoznak. Az optimális fokszámot az RMSECV értékek minimuma mutatja a 4.1.7. ábrán. Az RMSEP és RMSEC értékek megjelenítésével a túlillesztés mértékét is vizsgálni tudtam. Magas fokszám esetén az RMSEC értékek csökkenése mellett az RMSEP és RMSECV növekedését tapasztalhatjuk, mivel nagyobb fokszámú komponensek figyelembe vételével a véletlen hibákból következő ingadozásokra is ráillesztjük a görbét.



4.1.7. ábra: A pPLS modellekre megállapított RMSE értékek



4.1.8. ábra: Polinomiális PLS-sel végzett regresszió reziduum ábrája

A legjobb eredményeket a másodfokú polinom illesztése adta. A lineáris modellek 4.1.3. ábrán bemutatott reziduumainak parabolászerű elhelyezkedése ebben az esetben gyakorlatilag eltűnt (4.1.8. ábra). A modell tehát jól kezelte a vizsgált polimorf rendszer spektrumaiban előforduló nemlinearitást.

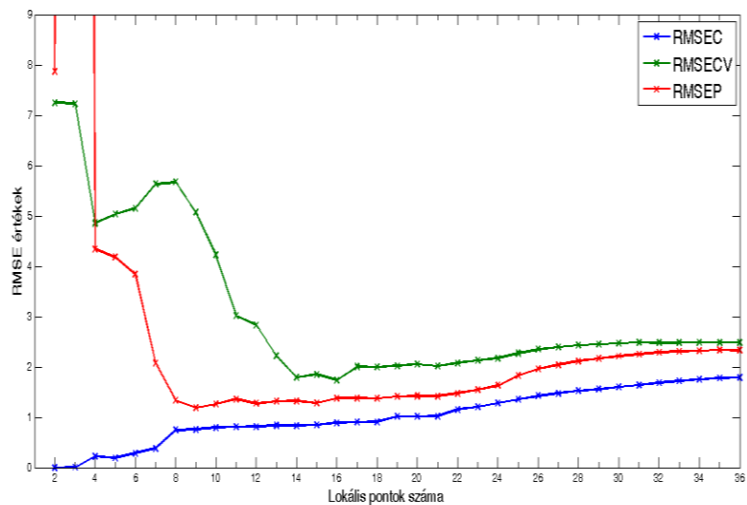
4.1.4.2. Helyileg súlyozott regresszióval végzett mennyiségi meghatározás

A helyileg súlyozott regresszió alkalmazása során a fő megválaszolandó kérdés az volt, hogy hány szomszédos pont figyelembe vételével illesszük a nemlineáris görbét a kalibráló pontokra. A pontok távolságát a főkomponens-térben vizsgáltam, a főkomponens-analízishez pedig a már bemutatott érvek alapján 2 látens változót használtam. A pontok súlyozásának módját nem változtattam, az alapértelmezett trikőbös súlyfüggvényt használtam, így csak a

gyakorlati alkalmazás szempontjából legfontosabb szomszédos pontok számának hatását vizsgáltam.

A különböző számú szomszédos pont használatával épített modellek hibaértékeit hasonlítottam össze az optimális érték megtalálásához (4.1.9. ábra). A lokális pontok számának minimuma 2, maximuma a kalibráló pontok száma (ekkor az összes pontot figyelembe veszi a regresszió során). Az ábrán nyomon követhető, hogy a keresztellenőrzési hiba körülbelül 15 szomszédos pontig csökken jelentősen, majd egy közel konstans szintre áll be, csakúgy mint a predikció hibája. Eközben a kalibráció hibája növekszik. 4 és 8 közötti lokális pont figyelembe vételével az RMSEP érték jelentősen megugrik. A megfelelő modellhez viszonylag sok pontot kell figyelembe venni. A hibaértékek

akkor voltak optimálisak, ha körülbelül a pontok felét szomszédosnak tekintettük (36 kalibráló pontból 16 szomszédos pont), és erre végeztük el lokálisan a súlyozott modell illesztését. Az ábrán is látható, hogy a módszer alkalmazásával kisebb RMSE értékek érhetők el, mintha az összes ponttal végezzük a regressziót (hagyományos PLS modell). A modell jóságát jellemző paramétereket a 4.1.4. táblázatban foglaltam össze.



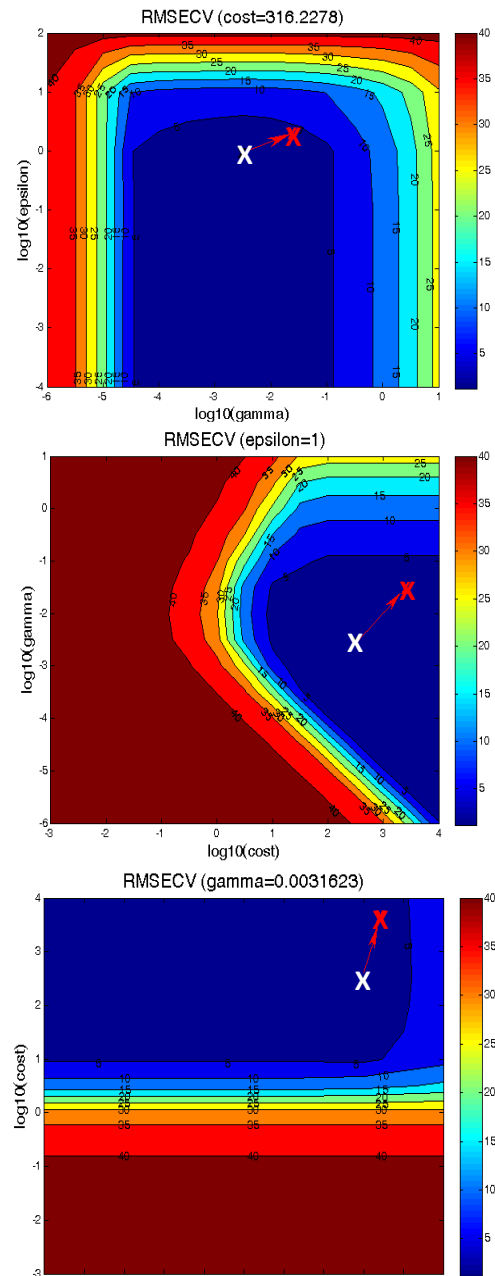
4.1.9. ábra: A különböző szomszédos pontok figyelembevételével felállított LWR modellek RMSE értékei (A 4-8 lokális pontot figyelembe vevő modellek esetén a RMSECV értékek kiugróan magasnak adódtak, így azok nem jelennek meg az ábrán.)

4.1.4.3. Támogatóvektorok módszerével végzett regresszió

A 3.6.5.3. pontban bemutatott támogatóvektorok módszerével, az adathalmaz magasabb dimenziójú térbe való vetítése által olyan nemlineáris problémák is kezelhetővé válhatnak, melyek megoldása a hagyományos módszerekkel nem lehetséges. A vetítést különböző kernelfüggvények segítségével valósíthatjuk meg, én a radiális bázisfüggvényt (Gauss görbékét) használtam munkám során.

A modellépítés során három paraméternek van döntő szerepe. Az ε paraméterrel beállíthatjuk, hogy mekkora legyen a regresszió során a tűréshatár, azaz hány pont legyen támogatóvektor. A C paraméterrel azt választjuk meg, hogy ezen pontok milyen arányossági tényezővel vegyenek részt a regresszióban. A γ paraméter a kernelfüggvény „kacsaringósságát” határozza meg, tehát a túlillesztés elkerülése miatt különösen fontos az értékének meghatározása.

A három paraméter optimális értékének megtalálásához a három paraméter RMSECV értékre gyakorolt hatását vizsgáltam a 4.1.10. ábra segítségével. Az ábrán azt láthatjuk, hogy az RMSECV értékeknek hol van minimuma két-két paraméter által kifeszített felületen, ha a harmadikat az optimális állandó értéken tartjuk. Az RMSECV értékek minimalizálásával azonban túl sok támogatóvektort használ a modell, mely túlillesztéshez vezet. Ezért úgy kell a paramétereket állítani, hogy egyben a támogatóvektorok száma is csökkenjen. Az ábrákról az is látszik, hogy ε értékétől kisebb mértékben függ az RMSECV érték. Azonban ha ezt a paramétert nagyobbra választottam, a támogatóvektorok számát sikerült csökkenteni. Ez alapján egy állandóan tartott, nagyobb ε értékhez kerestem meg γ és C paraméter optimumát. Az újbóli optimalás eredményeként a 33 támogatóvektor számát sikerült lecsökkenteni 7-re. Ennek az új optimumnak a helyét a 4.1.10. ábrán piros X jelzi. Ezeket a paramétereket használva, a modell előrejelző-képessége is javult, azaz RMSEP értéke csökkent (4.1.4. táblázat).



4.1.10. ábra: SVM optimalizálása. A fekete X az RMSECV minimalizálása alapján talált optimumot, a piros X a támogatóvektorok számának figyelembe vételével meghatározott optimumot jelzi.

4.1.5. A modellek összehasonlítása

4.1.5.1. A modellek összehasonlítása a jóságmutatók alapján

Az előző pontokban bemutatott, az egyes módszerekkel képzett legjobb modellek jóságmutatóit a 4.1.4. táblázatban foglalom össze. A táblázat alapján megállapítható, hogy a gyógyszeriparban leggyakrabban alkalmazott egyváltozós regresszió, a CLS, a PCR és a PLS modellezéssel elérhető pontosság javítható további többváltozós regressziós módszerek alkalmazásával.

4.1.4. táblázat: Az egyes regressziós módszerekkel képzett legjobb modellek jóságmutatói

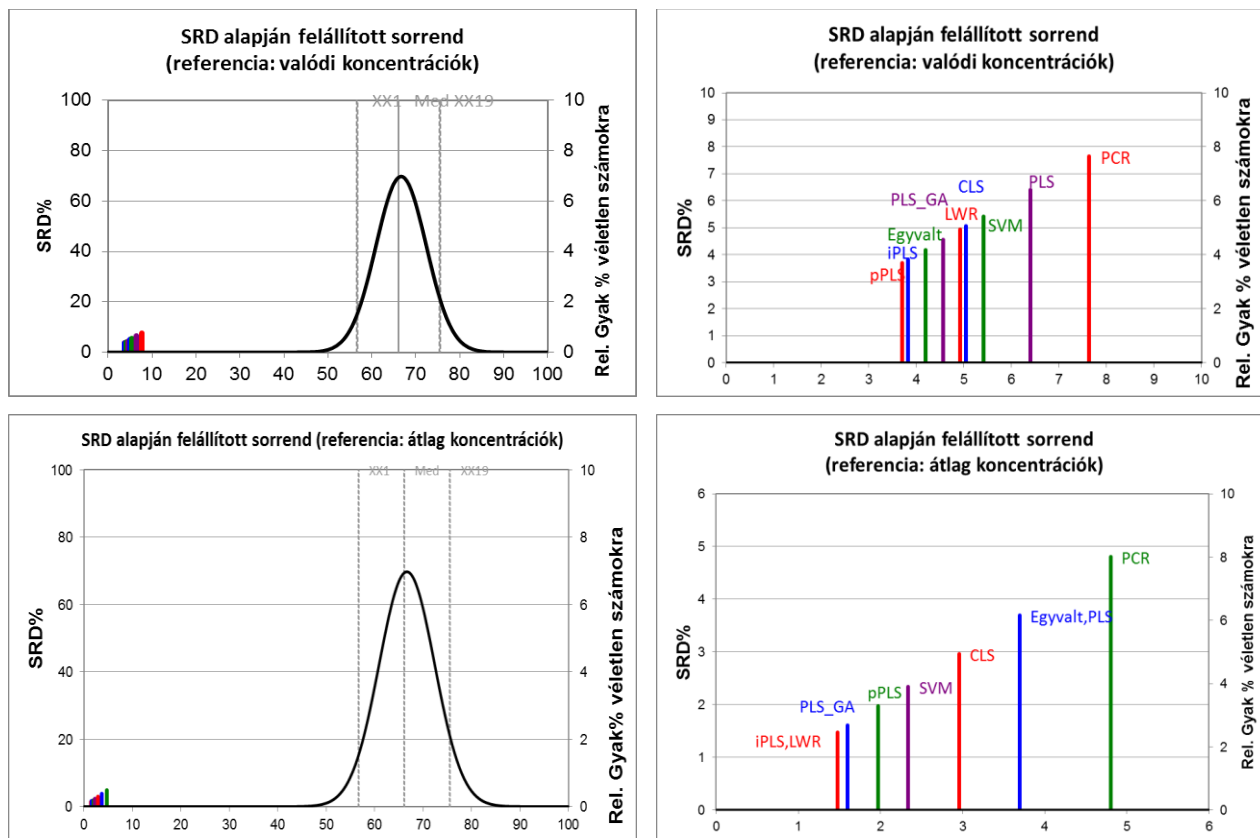
	Adat-előkészítés	RMSEC	RMSECV	RMSEP	R ² (cal)	R ² (CV)	R ² (pred)	Bias (cal)	Bias (CV)	Bias (pred)
Egy-vált.	alapv., normálás	3,20	-	3,16	0,9941	-	0,9940	-0,01	-	-0,74
CLS	alapv., normálás	3,48	-	3,63	0,9931	-	0,9921	-0,02	-	-0,03
PCR	alapv., normálás, centrálás, SNV	1,82	4,00	1,88	0,9981	0,9912	0,9984	0	-0,83	-0,90
PLS		1,68	2,16	1,91	0,9984	0,9973	0,9983	0	0,096	-0,92
iPLS		1,14	1,30	1,20	0,9993	0,9990	0,9994	0	0,0076	-0,60
PLS-GA		1,01	1,33	1,19	0,9994	0,9990	0,9994	0	-0,02	-0,50
pPLS	alapv., normálás, centrálás,	1,06	1,67	1,24	0,9994	0,9985	0,9993	-	-	-
LWR	alapv., normálás,	0,89	1,74	1,38	0,9996	0,9984	0,9992	-0,035	-0,56	-0,78
SVM	alapv., normálás, centrálás, SNV	1,16	1,54	1,13	0,9992	0,9987	0,9994	-0,07	-0,18	-0,53

Míg az egyváltozós regresszió illetve a klasszikus legkisebb négyzetek módszere 3%-nál nagyobb hibával becsüli az egyes koncentrációkat, nemlineáris modellek és változókiválasztás alkalmazásával ez lecsökkenthető 1%-ot megközelítő értékre. Ez jelentős javulást jelent, tekintve, hogy a polimorfok mennyiségi meghatározására főleg kis koncentrációknál van szükség. Látható

az is, hogy az adatelőkészítések közül főleg a centrálás és a sorstandardizálás alkalmazása javasolt.

4.1.5.2. Modellek összehasonlítása a Rangszámkülönbségek abszolútérték-összegének módszere (SRD) alapján

A modellek összehasonlításának objektív lehetősége a Rangszámkülönbségek abszolútérték-összegének módszere (SRD). A referencia jelen esetben lehet a bemért koncentráció, illetve a modellekből számolt koncentrációk átlaga is.



4.1.11. ábra: Az SRD módszerrel felállított sorrendek a kétféle referencia-sorrend esetén

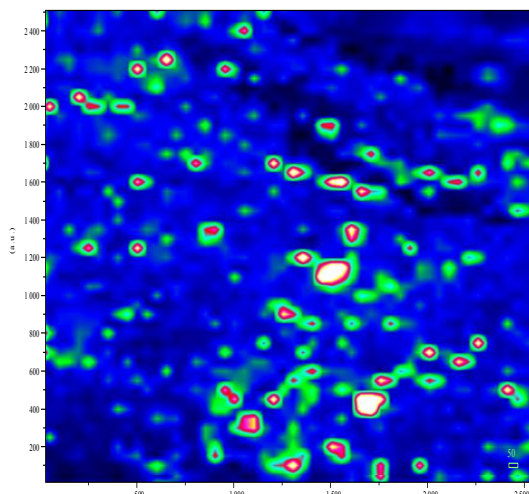
A referenciákra mindkét lehetőséget megvizsgáltam, mert ezáltal nyílik lehetőségünk arra, hogy a referenciáktól függetlenül megállapításokat vonjunk le. A bemért koncentrációkat referenciaként használva a módszer azt jelzi érzékenyen, ha a modell a közeli koncentrációtartományokat nem tudja megfelelően elkülöníteni. Így a modelleket az RMSE értékek alapján való összehasonlítástól különböző szempontból tudjuk tanulmányozni, az SRD eredményei leginkább a reziduumok alapján történő vizsgálatnak feleltethetjük meg. A kapott

sorrendeket a 4.1.11. ábra mutatja be. Az SRD értékeket itt százalékosan ábrázoltam. Ekkor a 100%-ot a referenciához képest létrehozott teljesen fordított sorrend adja.

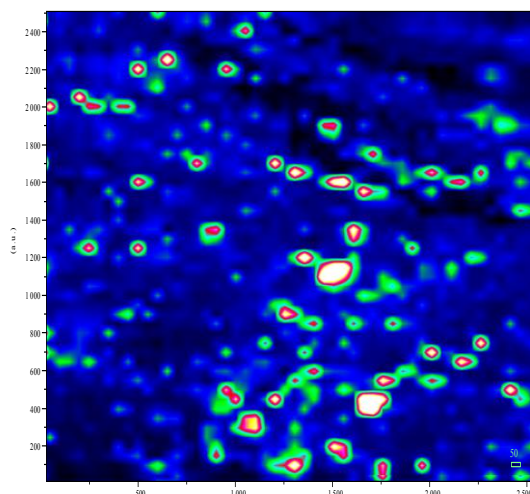
A véletlen számok Gauss-görbét követő eloszlását is feltüntettem a diagramokon. A baloldali diagramokon látható, hogy az összes modell jelentősen eltér attól, mintha véletlen számokat írnánk becslésnek. A láthatóság érdekében a releváns tartományokra ránagyítottam. Kitűnik, hogyha az átlagkoncentrációkat használtam referenciának, akkor a változókiválasztással támogatott PLS regresszió valamint a nemlineáris modellek szerepeltek a legjobban, míg a referenciamódszerek (egyváltozós regresszió, CLS) elmaradtak ezektől. A valós koncentrációk alapján felállított sorrend is azt mutatja, hogy a gyógyszeripari gyakorlatban leggyakrabban használt PCR és PLS módszerek képesek a legkevésbé elkülöníteni az egyes koncentrációtartományokat, míg a változókiválasztással segített és a nemlineáris módszerek ez alapján is pontosnak bizonyultak.

4.1.5.3. *Modellek használata új térkép mennyiségi elemzésére*

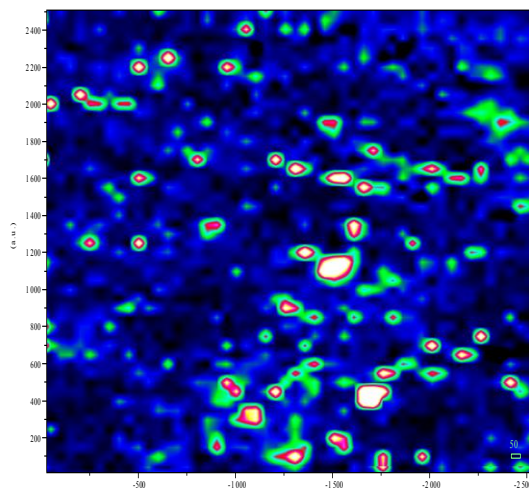
A kémiai térképezéssel és megfelelő regressziós modellek használatával elérhetjük, hogy bármely a gyártásból származó tételnek a pontos összetételén kívül a komponensek pontos eloszlástérképeit is megkapjuk. Ez a technológusok számára további információkat hozhat. A bemutatott modellépítések során a legfőbb szempont a modellek megfelelő előrejelző-képességének elérése volt. A jövőbeni mintákon való használhatóság jellemzésére az RMSECV és RMSEP értékeket használtam, a modellek optimalizálásával pedig igyekeztem elkerülni a túlillesztést. Az így felépített modellek teljesítményét vizsgáltam a modellépítés során egyáltalán nem használt térkép koncentrációprofiljának elemzésével. Ehhez az 5% Form I-et és 95% Form II polimorfot tartalmazó keverék 101×101 képpontból álló Raman-térképét vizsgáltam.



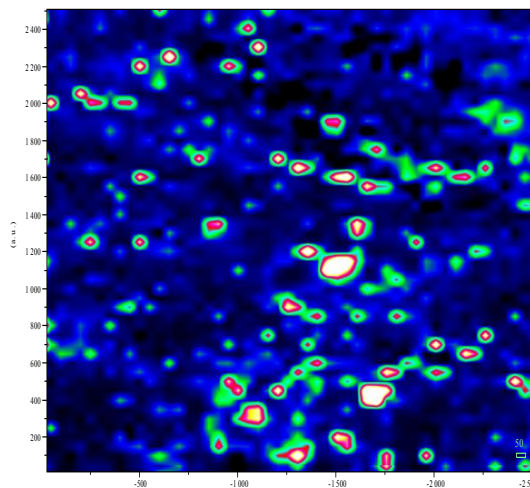
CLS modell
Átlag= 7,06 %
 (Minimum= 0,07 Maximum= 76,16)
 Szórás= 6,92



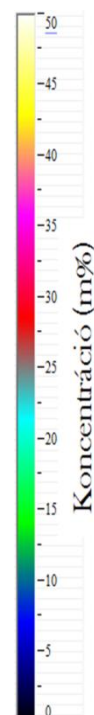
PLS modell
Átlag= 5,76 %
 (Minimum= -0,89 Maximum= 79,51)
 Szórás= 7,67



iPLS modell
Átlag= 5,45 %
 (Minimum= -8,87 Maximum= 90,45)
 Szórás= 8,48



LWR modell
Átlag= 5,26 %
 (Minimum= -6,89 Maximum= 79,11)
 Szórás= 8,34



4.1.12. ábra: Négy különböző módszer által becsült koncentrációeloszlás-térképnek egy kinagyított részlete illetve a koncentráció-térképekből számolt statisztikai jellemzők

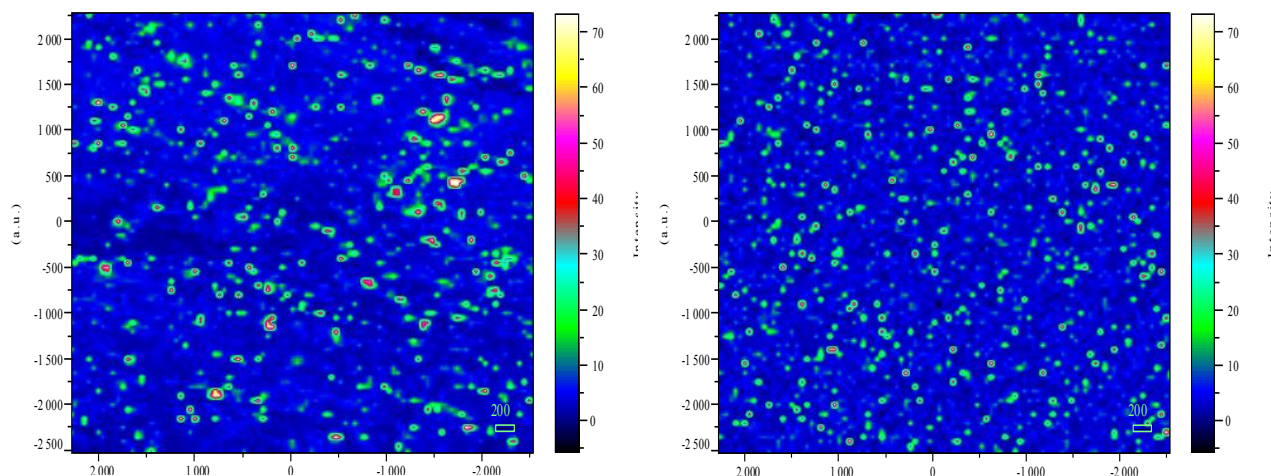
A 4.1.12. ábrán látható négy modell alapján számolt koncentrációtérképnek egy kinagyított részlete (a teljes térképeken kevésbé figyelhetők meg a különbségek). A koncentráció-térképeken viszonylag kevés eltérés látható, azonban a koncentrációkból számolt

statisztikai jellemzők alapján megállapítható, hogy a nemlineáris LWR (5,25%) és SVM (4,76 %) adja vissza legpontosabban a névleges 5% Form I tartalmat. Az SVM a modellek közül egyedülálló módon alábecsüli ezt az értéket. Fontos tapasztalat, hogy a PLS-GA a kis RMSE értékek ellenére az új térképet kevésbé pontosan becsli, mint a PLS (5,87% átlagkoncentráció), tehát a modell túlillesztett.

4.1.6. A térképméret homogenításra gyakorolt hatásának vizsgálata

Azt, hogy a mennyiségi meghatározáshoz használt térképméret megfelelően reprezentatív-e, az előző pontban már bemutatott 101 ×101 képpontból álló térkép makropixel-elemzésével vizsgáltam. Amennyiben a mintaelőkészítést megfelelően végeztük, ez a térképméret már biztosan megfelelő nagyságú a minta koncentrációjának meghatározásához. Ezt a feltevést az előző pontban bemutatott eredmények támasztják alá.

Munkám során a PLS modellezéssel nyert koncentráció-térkép egy 96×96-os részletét vizsgáltam (ezt a méretet viszonylag sok méretű makropixelre lehet bontani). A térképet először randomizáltam (4.1.13. ábra), majd a térképeket 2×2, 3×3, 4×4, 6×6, 8×8, 12×12, 16×16, 24×24, 32×32 és 48×48-as méretű makropixelekre bontottam és mindegyik esetben kiszámoltam a Poole-indexet a 3.7. fejezetben bemutatottak szerint.

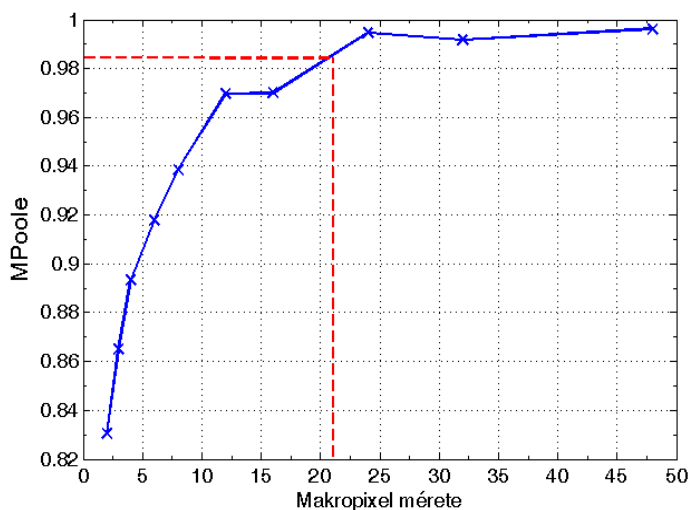


4.1.13. ábra: A PLS modellezéssel nyert koncentráció-térkép és annak randomizált térképe

Ezeket az értékeket a makropixel-méret függvényében ábrázoltam (4.1.14. ábra). A diagramon jól látszik, hogy a térképméret növelésével a homogenitás is növelhető, míg meg nem közelítjük a homogén térképet jelentő 1 értéket.

Megállapítottam, hogy a polimorf rendszer kvantitatív jellemzéséhez alkalmazott 21×21 képpontot tartalmazó térkép Poole-index szerinti homogenitása fokozható lenne nagyobb térkép felvételével, azonban már ennél a térképméretnél is megfelelőnek tekinthető. Nagyobb térkép felvétele a mérési idő növekedéséhez vezetne, mely különösen a Raman-térképezés folyamatos technológiákban való alkalmazása esetén

jelentős hátrány. Ugyanakkor a bemutatott módszer alkalmazásával lehetőség nyílik arra, hogy a kvantitatív meghatározás érdekében felvett térkép nagyságát tudatosan, a mérési idő és a reprezentatív „mintavétel” együttes figyelembevételével határozzuk meg.

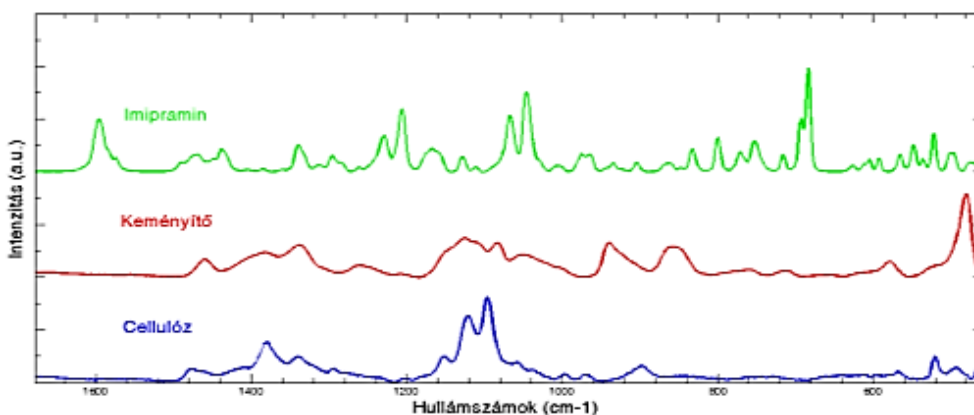


4.1.14. ábra: A homogenitás növekedése a térképméret növelésével

Egy gyógyszeripari szempontból is releváns polimorf rendszer példáján tanulmányoztam a keverékek mennyiségi elemzésének lehetőségét az egyes többváltozós regressziós módszerek alkalmazásával. A kvantitatív meghatározás szempontjából ez egy viszonylag egyszerű, kétkomponensű rendszer volt, mely lehetőséget adott arra, hogy megvizsgáljam a különböző kemometriai módszerek használatának fő szempontjait. A gyakorlat során azonban ennél összetettebb rendszerek pontos elemzésére is szükség van, mely új kérdéseket, problémákat vet fel a modellek építése és használata szempontjából. Dolgozatom további részében a kétkomponensű rendszer vizsgálata során nyert tapasztalatokra építve mutatom be egy háromkomponensű, egy hatóanyagot és két segédanyagot tartalmazó rendszer kvantitatív meghatározását. Munkám során annak felderítésére helyeztem a hangsúlyt, hogy milyen új kihívásokat hoz elő több komponens egyidejű mennyiségi elemzése illetve milyen eszközökből választhatunk ezen problémák kezelésére.

4.2. Hatóanyagot és segédanyagokat tartalmazó keverékek mennyiségi elemzése

A hatóanyagot és segédanyagokat tartalmazó rendszer mennyiségi elemzése során olyan modell megalkotására törekedtem, mely modellel egyszerre nyílik lehetőség mindhárom komponens pontos meghatározására. Ez több szempontból is kihívást jelentő feladat. A keverék komponensei közül a keményítő és cellulóz szerkezete nagyon hasonló, ez Raman-spektrumuk hasonlóságaiban is megnyilvánul, egymás melletti kiértékelésük akadályokba ütközhet. Másrészt Raman-aktivitásuk jelentősen kisebb az imipraminéhoz képest: míg az imipramin Raman-spektruma sok és éles csúcsot tartalmaz, a segédanyagokat széles, kevésbé intenzív sávok jellemzik (4.2.1. ábra). Ebből arra következtethetünk, hogy a segédanyagokkal szemben az imipramin pontosabb leírására lesznek alkalmasak a kvantitatív modellek. Továbbá nem hagyhatjuk figyelmen kívül, hogy a segédanyagok makromolekulás anyagok, több ezer g/mol-os, molekulánként változó móltömeggel, szemben az imipramin 280,41 g/mol-os tömegével. A Raman-spektrometria alapvetően a jelet okozó kötések számával, ezáltal a móltörttel arányos jelet szolgáltat. Ez a tény a tömegszázalékok szerinti kalibrálásban okozhat problémát (például nemlinearitás megjelenését).



4.2.1. ábra: A keveréket alkotó komponensek alapvonal-korrigált és normált referenciaspektrumai

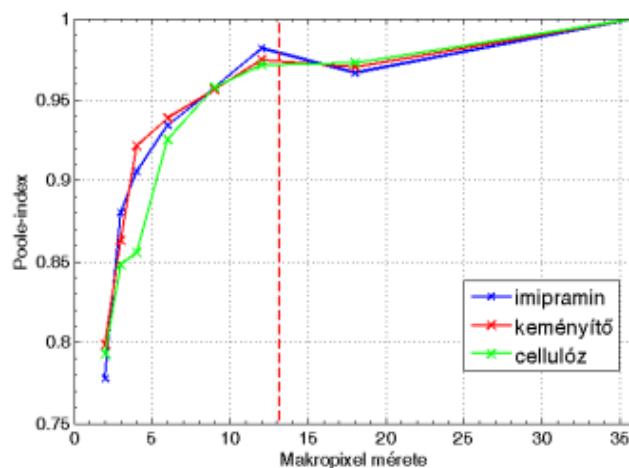
4.2.1. Az optimális térképméret meghatározása

A megfelelő kvantitatív elemzéséhez szükséges térképméret meghatározása többkomponensű rendszer esetén nagy körütekintést igényel. A felvett térkép jól kiválasztott nagyságával tudjuk biztosítani, hogy a térképpontokon mindegyik komponensre megfelelően reprezentatív mennyiségeket kapjunk. Az előző fejezetben, a carvedilol térkép homogenitás vizsgálatára is

alkalmazott makropixel-elemzés használható módszerek tűnt a megfelelő térképméret megállapításához, ezért a háromkomponensű keverék-rendszerről felvett térképek méretét is ezzel a módszerrel határoztam meg.

Ennek érdekében az M9-es (50% keményítő, 25% cellulóz és 25% imipramin tartalmú) mintáról 49×49 képpontból álló térképet vettem fel, majd CLS modellezés segítségével határoztam meg a komponensek koncentráció-eloszlását. A koncentráció-térképeknek egy 36×36 -os nagyságú részletét használtam fel. A koncentráció-térképeket randomizáltam majd a valódi és a randomizált térképet is makropixelexre bontottam, melyek 2×2 , 3×3 , 4×4 , 6×6 , 9×9 , 12×12 , 18×18 , 36×36 méretűek voltak. A 3.7. pontban bemutatott módon kiszámolt Poole-indexek a makropixelex függvényében ábrázolva láthatók a 4.2.2. ábrán.

Az ábráról kitűnik, hogy az egyes komponensek homogenitása közel azonos, a cellulóz marad el kissé az imipraminhoz és keményítőhöz képest. A 18×18 méretű makropixelnél tapasztalható enyhe homogenitáscsökkenés valószínűleg a randomizáláskor előforduló véletlen hatások eredménye. Ez fakadhat abból, hogy például egyes nagyobb koncentrációjú pontok egy helyen sokasodnak a randomizált térképen, aminek bekövetkezésére van egy kevés valószínűség. A 36×36 -os méretnél a homogenitás eléri az 1-et, de már a 9×9 -es makropixel is megfelelően homogénnek tekinthető, 0,95-nél nagyobb Poole-indexszel. Ezen megfontolások alapján illetve a mérési időt is figyelembe véve, 13×13 képpontból álló térképeket vettem fel. Ez a térképméret már megfelelően leírja a mintára jellemző átlagkoncentrációt.



4.2.2. ábra: Makropixel-elemzés eredménye

4.2.2. Adatelőkészítési módszerek vizsgálata

A térképek spektrumain a dióda lézer használatának ellenére is viszonylag erős fluoreszcens háttér figyelhető meg, elsősorban a cellulóz spektrumai esetében. A sikeres mennyiségi meghatározáshoz tehát szükség van alapvonal- korrekcióra. Ahhoz, hogy az

alapvonal- illesztések eltérése ne legyen befolyásoló tényező, minden spektrumra ugyanolyan alapvonal-korrekciónak célszerű alkalmazni. A polimorf keverékek esetében alkalmazott szakaszosan lineáris függvény illesztését ebben az esetben több okból sem tudtam használni. Egyrészt a komponenseknek erősen eltért egymástól az alapvonala. Másrészt sokszor a segédanyagoknak ott jelentkeztek széles sávjai, ahol az imipraminnak szükséges lett volna alapvonali pontokat megadni. Végül nem találtam mindhárom komponensre ugyanolyan jól illeszkedő szakaszos függvényt. Ennek a problémának megoldására az Eilers-féle, aszimmetrikus legkisebb négyzetek módszere szerinti alapvonal-illesztést [86] alkalmaztam. Minden térkép minden spektrumát az empirikusan megállapított $p=0,001$ és $\lambda=10^5$ paraméterekkel illesztett alapvonallal korrigáltam.

A fókuszálásból adódó intenzitásingadozások kiküszöbölésére normálásra van szükség, azonban alkalmazása ebben az esetben jelentős torzítást visz a rendszerbe, mert az imipramin spektrumainak jóval nagyobb az intenzitása, mint a keményítőnek illetve a cellulóznak. Annak eldöntésére, hogy ennek ellenére érdemes-e normálni a térképeket, megvizsgáltam egy CLS modellt görbe alatti területre normálás alkalmazásával és anélkül. Azt tapasztaltam, hogy a normálás használatával az imipramin meghatározási hibája néhány százalékkal nőtt, hiszen ennek spektrumát torzította legjobban ez az adatelőkészítés. Ugyanakkor a keményítő és cellulóz esetén a normálás jobb meghatározást tesz lehetővé (4.2.1. táblázat).

4.2.1. táblázat: A normálás hatásának vizsgálata CLS modellen

	CLS modell alapvonal-korrekciónal			CLS modell alapvonal-korrekciónal és normálással		
	keményítő	cellulóz	imipramin	keményítő	cellulóz	imipramin
RMSEC	8,37	10,52	6,41	6,82	5,76	10,04
RMSECV	10,32	13,20	7,93	8,50	7,11	12,60
RMSEP	8,81	11,42	10,35	6,20	6,71	12,51

Mivel olyan modellekre törekedtem, melyek mindhárom komponens mennyiségi meghatározását lehetővé teszik, úgy döntöttem, hogy a normált spektrumokkal végzem a modellépítéseket. A vizsgált rendszerben a segédanyagok mennyiségi meghatározása volt a nagyobb kihívást jelentő feladat, míg az imipraminra pontosabb eredményeket vártunk a sok szelektív csúcsnak köszönhetően. A CLS modell helyett is más kemometria módszerekhez kellett folyamodnom, hogy ezeket a pontatlanságokat és a torzításokat csökkentsem.

Hasonlóan a carvedilol rendszer - előző fejezetben ismertetett – vizsgálatához, ebben az esetben is teszteltem egyéb adatelőkészítési módszereket (centrálás, autoskálázás, MSC, sorstandardizálás), azonban a centráláson kívül ezek a módszerek nem adtak egyik modell esetében sem pontosabb eredményeket.

4.2.3. A modellek validálása

A kvantitatív modellek jóságának jellemzésére a polimorfok esetén már megszokott közepes négyzetes hibákat (RMSEC, RMSECV és RMSEP), a determinációs együtthatókat (R^2), a tengelymetszeteket valamint a reziduum-ábrákat alkalmaztam. Minden értéket külön-külön számoltam a három komponensre, a modellek megfelelőségét ezáltal mindegyik összetevőre vizsgálhattam.

A modellek felépítéséhez a minták alapvonal-korrigált majd normált térképeinek átlagát használtam. A modellek belső validálását egy elem kihagyásos keresztellenőrzéssel végeztem azon módszerek esetén, melyek ezt lehetővé tették. A mintasorozatot kalibráló- és validáló halmazra osztottam.

A mintasorozat felosztása kalibráló és validáló halmazra

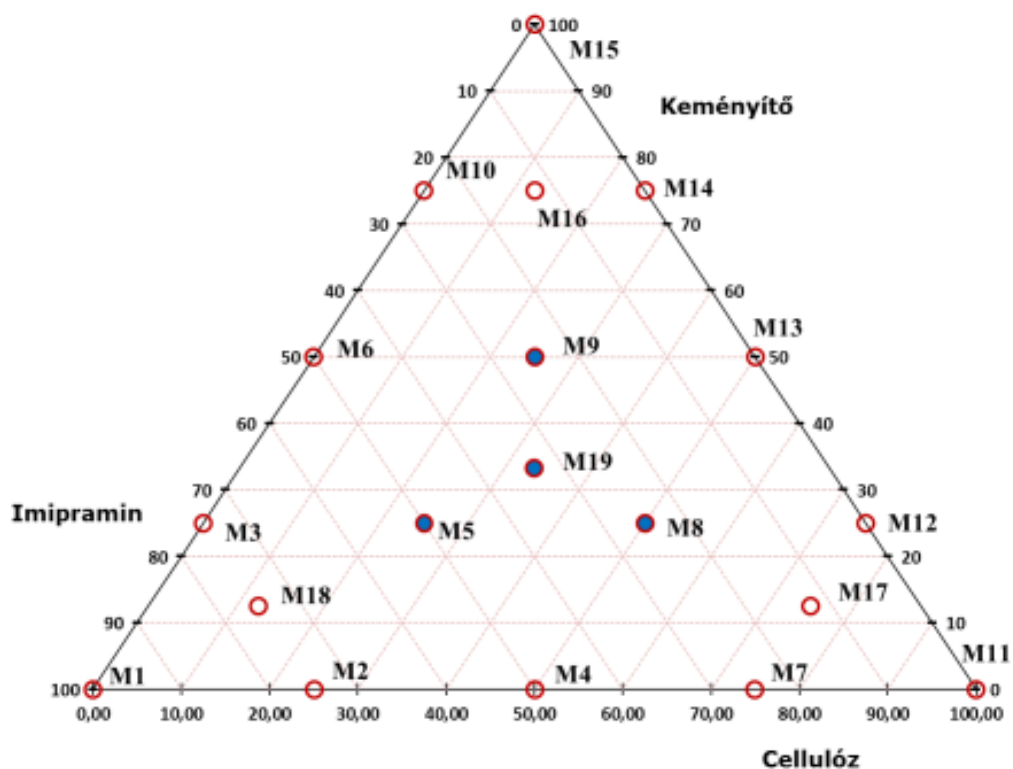
A kétkomponensű polimorf rendszer esetén a mintasorozat felosztása kalibráló illetve validáló halmazra viszonylag egyszerű feladat volt, csupán arra kellett figyelemmel lenni, hogy a két halmaz között a pontokat a teljes koncentrációtartományban nagyjából egyenletesen osszuk el. Több komponens esetén azonban ennek biztosítása nehezebb feladat. Az általam vizsgált háromkomponensű rendszerben viszonylag kevés, 19 pontot kellett úgy elosztani, hogy a kalibráló halmazba kerülő pontok mindegyik komponenst egyaránt jól le tudják írni.

Ezek alapján azt feltételeztem, hogy a kalibráló és validáló halmazra történő felosztásnak nagy hatása lehet a kvantitatív modellek eredményére. Ezért megvizsgáltam, hogy különböző kalibráló mintákkal épített modellek között észlelhető-e különbség illetve milyen felosztást érdemes választani a jó eredmények érdekében.

Teszteket végeztem el 80:20 illetve 60:40 arányú felosztással, az első felosztás szerint 15 minta került a kalibráló halmazba és 4 a validálóba, a második esetben pedig 8 mintával validáltam. Olyan felosztásokat kerestem, melyek valamilyen szimmetria szerint helyezkednek el a minták által kifeszített háromszögdiagramon (4.2.3. ábra) valamint olyan modelleket is vizsgáltam, melyeknél a felosztás „random” módon, véletlen mintakiválasztással történt. A

vizsgált felosztásokat a 4.2.2. táblázatban foglaltam össze. Értelmeszerűen az 1-es számú felosztás a 80:20, míg a 2-es számmal kezdődők a 60:40 arányt követték. A validáló halmazba került minták neveit mindegyik esetén feltüntettem.

A különböző módon felosztott halmazokkal, alapvonal-korrekción, normálás és átlagra centrálás után 3 látens változót használó PLS modelleket építettem, és a modellek összehasonlítására a 3-3 RMSEC, RMSECV és RMSEP értéket alkalmaztam.

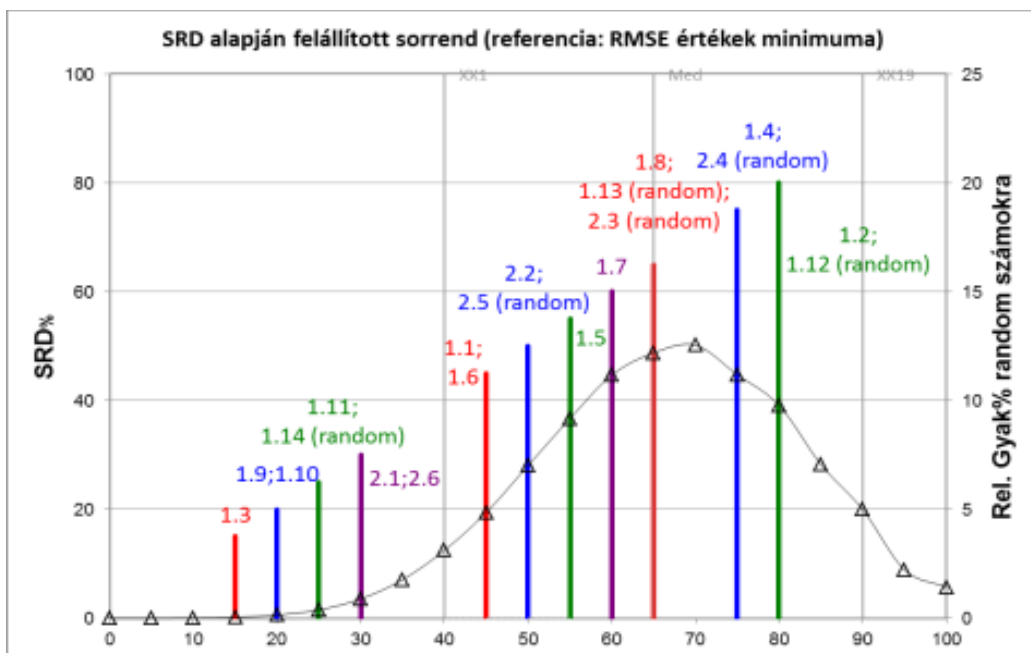


4.2.3. ábra: A mintasorozat elemeinek elhelyezkedése háromszögdiagramon. A kék pontok a vizsgálat eredményeként legjobbnak tekinthető modell validáló pontjait jelzik.

4.2.2. táblázat: A különböző kalibráló-validáló halmazokra való felosztások

Felosztás száma	Validáló halmazba kerülő minták neve	Felosztás száma	Validáló halmazba kerülő minták neve
1.1.	M16 M17 M18 M19	1.11.	M10 M14 M2 M7
1.2.	M4 M6 M13 M19	1.12. (random)	M1 M6 M12 M18
1.3.	M5 M8 M9 M19	1.13. (random)	M6 M10 M12 M14
1.4.	M1 M11 M15 M19	1.14. (random)	M2 M5 M13 M18
1.5.	M15 M16 M9 M19	2.1	M16 M17 M18 M19 M5 M8 M9 M4
1.6.	M15 M16 M9 M4	2.2.	M3 M6 M10 M12 M13 M14 M2 M7
1.7.	M1 M18 M5 M13	2.3. (random)	M1 M2 M5 M8 M11 M13 M16 M18
1.8.	M11 M17 M8 M6	2.4. (random)	M1 M3 M6 M9 M12 M14 M17 M19
1.9.	M3 M7 M10 M12	2.5. (random)	M3 M4 M7 M11 M12 M13 M14 M17
1.10.	M2 M3 M7 M10	2.6.	M5 M8 M9 M19 M3 M7 M10 M12

A kilenc RMSE érték alapján a modelleket a Rangszámkülönbségek abszolútérték-összegének módszere (SRD) segítségével rangsoroltam, az RMSE értékek minimumát használva referenciaként. Az eredményül kapott 4.2.4. ábrán megfigyelhető, hogy a vizsgált 20 felosztás közül csupán hét az, ami különbözik a véletlen számokkal végzett rangsorolástól (a Gauss-görbe előtt van). Az is észrevehető, hogy a random felosztások, egy kivételével, magasabb rangszámokat kaptak, mint a valamilyen szimmetriaelv szerint elrendeződő pontokat tartalmazók. A megfigyeléseim alapján kijelenthetem, hogy a felosztás során célszerű a szimmetrikusan kiválasztani a validáló elemeket. Erre világít rá az is, hogy az 1.9., 1.10. és 1.11. számú felosztások ugyanolyan szimmetriaelvet követtek, és a rangsorolásban ezek azonos illetve egymás melletti helyen végeztek. Szintén fontos tapasztalat, hogy a 60:40 felosztású minták rosszabbnak bizonyultak. Ez összhangban van azzal a feltevessel, hogy a három komponens megfelelő leírása viszonylag sok pontot igényel. A rangsorolás alapján az 1.3. számú felosztást találtam a legjobb modellnek. Ennek validáló pontjait a 4.2.3. ábrán kék pontokkal jelzem. Ezt a felosztást használtam a továbbiakban minden egyéb kiértékelésben.



4.2.4. ábra: A különböző kalibrációs halmazokkal nyert modellek rangsorolása SRD alapján

A bemutatott eredmények alátámasztják azt a feltételezést, hogy a kalibráló pontok megválasztása jelentősen befolyásolhatja a modellépítések eredményeit. Ennek a hatásnak részletesebb felderítése további vizsgálatokat igényel, melyet lehetőség szerint kísérletterv alapján szeretnék megvalósítani.

4.2.4. Referenciamódszerekkel elért eredmények

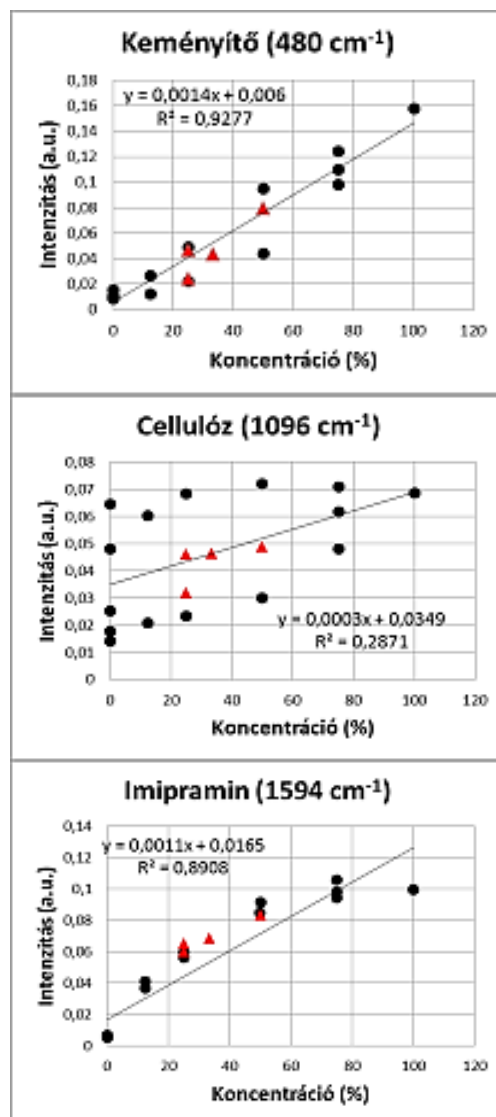
4.2.4.1. Egyváltozós regresszió

A vizsgált háromkomponensű rendszer egyváltozós kiértékelése problémákba ütközik amiatt, hogy a segédanyagok, főleg a cellulóz spektrumában nem található nagy szelektivitású csúcs. Szintén figyelembe kell venni azt, hogy egyváltozós regresszió során nem tudunk több függő változót (azaz a három komponens koncentrációértékét) egyszerre kezelni. A felsorolt korlátozó tényezők ellenére elvégeztem az egyváltozós regressziót, mivel az ezzel nyert eredmények jól mutathatják, hogy milyen tényezőkre kell figyelemmel lenni a többváltozós modellek alkalmazása során.

A kiértékeléshez az imipramin 1594 cm^{-1} , a keményítő 480 cm^{-1} és a cellulóz 1096 cm^{-1} -nél lévő hullámszámát választottam. Ahogy az a 4.2.1. ábrán is megfigyelhetjük, az első két komponensre ezek viszonylag szelektív csúcsok, a cellulóz esetén azonban az 1096 cm^{-1} kevésbé különül el a többi komponens spektrumától.

A regressziós egyeneseket a 4.2.5. ábrán mutatom be, az RMSE és R^2 értékek pedig a 4.2.3. táblázatban találhatók.

Az eredményekből kitűnik, hogy a cellulóz mennyiségi meghatározására a módszer nem alkalmas. A regresszió az imipraminra, szelektív csúcsának ellenére, viszonylag nagy hibaértéket adott (12%-os kalibrációs hiba), melynek oka a korábban részletezett normálás miatti torzítás lehet, mely különösen erősen jelentkezik, ha csak egy hullámszámon leolvasható intenzitásértékkel becsülünk koncentrációt. Erre utal az imipramin regressziós görbéjén tapasztalt erős nemlinearitás is. A módszerrel a keményítő koncentrációja határozható meg legkisebb, 10% alatti hibaértékkel, azonban ez a pontosság sem felel meg a gyógyszeripari elvárásoknak.



4.2.5. ábra: Egyváltozós regressziós egyenesek

4.2.4.2. Klasszikus legkisebb négyzetek módszere és főkomponens-analízisen alapuló regresszió

A CLS regressziót kalibráló sor felhasználásával végeztem el, alapvonal-korrigált és normált spektrumokkal. A módszer az egyváltozós regresszióhoz képest mindhárom komponens esetén csökkentette a meghatározási hibát. Legnagyobb mértékű javulást a cellulóz esetén tapasztaltam, RMSE értékei lecsökkentek 10% alá, így azok összevethetőek lettek a keményítő hibaértékeivel. Azonban az imipramin meghatározási pontossága jelentősen elmaradt a segédanyagoknál elérhetőtől (10-12%-os RMSE érték), mely oka még ebben az esetben is a

normálással bevitt torzítás lehet, melyet CLS modellezéssel nem tudtam megfelelően kezelni. (4.2.3. táblázat)

A főkomponens-regresszió és részleges legkisebb négyzetek módszerével is történt modellezés. A modellekben felhasznált főkomponensek/ látens változó számát mindhárom komponens RMSE értékeit figyelembe véve határoztam meg azért, hogy semelyik komponens mennyiségi meghatározására vonatkozóan ne legyen a modell túlillesztett. A legkisebb RMSE értékekre törekedtem, ahogy eddig is. Az RMSE értékeknek 2 illetve 4 látens változó között ingadozott a minimum értékük. Ez alapján a 3 látens változót alkalmazó modell építése mellett döntöttem, mely összhangban van a rendszer komponensszámával is. A spektrumok átlagra centrálásával mind a PCR, mind a PLS modell jelentősen pontosabb lett. Ennek az adatelőkészítésnek köszönhetően kiküszöböltem azt a hibát is, melyet a normálással vittem az adatokba. A modellt már az imipramin meghatározására is megfelelően lehetett alkalmazni, így már mindegyik komponensre vonatkoztatott hiba 10% alá csökkent (4.2.3. táblázat)

4.2.5. Változókiválasztással támogatott PLS regresszió

Az előző pontban bemutatott PLS modell megfelelően bizonyult mindhárom komponens közel azonos hibával történő becslésére, azonban ennek a hibának további csökkentésére a gyógyszeriparban és a készítmények minőségbiztosításában nagyon gyakran szükség lehet. Ennek érdekében megvizsgáltam a polimorfok elemzése során már hasznosnak bizonyult változókiválasztási módszereket.

Megvizsgáltam, hogy az intervallum-PLS és a genetikus algoritmusokkal támogatott PLS regresszióval milyen eredményeket nyerhetünk. Az intervallum-PLS alkalmazása során egyaránt tanulmányoztam a *forward* és *reverse* futási formát is, azonban nem tudtam egyértelmű következtetések levonni arra vonatkozóan, hogy a modellépítés eredményét befolyásolja-e a futtatás módja. Mind az iPLS, mind a PLS-GA esetén azt tapasztaltam, hogy a modelljóságot leginkább befolyásoló paraméter a változóablakok mérete. 50, 25, 10 és 5 változót tartalmazó ablakokkal végeztem el a modellezéseket. A tapasztalatok szerint a kis változóablakok esetén a módszerek viszonylag sok ablakot választottak ki és ezáltal a modellek RMSE értékei csökkentek, ugyanakkor kisebb mértékű túlillesztés lépett fel, mint a polimorf rendszeren végzett változókiválasztás során, ugyanekkora ablakmérettel dolgozva.

A változókiválasztással támogatott modellek a három komponens meghatározási hibáját különbözőképpen befolyásolják: egyes modellek esetén egy-egy komponenst jellemző RMSE értékek nőttek a PLS modellhez képest, míg más komponens esetén jelentős javulást tapasztaltam. A mindhárom komponensre kapott RMSEC, RMSECV és RMSEP értékeket éppen ezért egyszerre viszonyítottam egymáshoz. Az összehasonlítást oly módon végeztem el, hogy a kilenc RMSE érték és a kilenc determinációs együttható szerint rangsoroltam a modelleket. Egy-egy modelljóságot leíró paraméterre mindig meg tudtam állapítani a modellek közti sorrendet (a legkisebb RMSE illetve a legnagyobb R^2 értékű modell kapta a legkisebb rangszámot). A kapott rangszámokból minden modellre kiszámoltam az átlagos rangszámot. Értelemszerűen a legkisebb átlagos rangszámú modellt választottam a legjobbnak. A legjobbak az átlagos rangszám szerinti sorrendbe állítás alapján az 5 és 10 változót tartalmazó változóablakot használó iPLS modellek lettek, illetve hasonlóan jól szerepeltek az ugyanekkora ablakméretet alkalmazó PLS-GA modellek is. Ez a sorba rendezési eljárás azonban a túlillesztést nem veszi figyelembe. Ahhoz tehát, hogy jó döntést hozzunk, a túlillesztési tényezőkre is tekintettel kellett lennünk.

A megfelelő változókiválasztással a meghatározások hibáit 5% alá sikerült csökkenteni. Amennyiben a gyakorlati probléma nem igényli mindhárom komponens pontos meghatározását, található olyan változókiválasztással támogatott PLS modell is, mellyel akár 2-3%-ra is lecsökkenthető a kívánt komponens meghatározási hibája (a másik két komponens magasabb hibaértéke mellett).

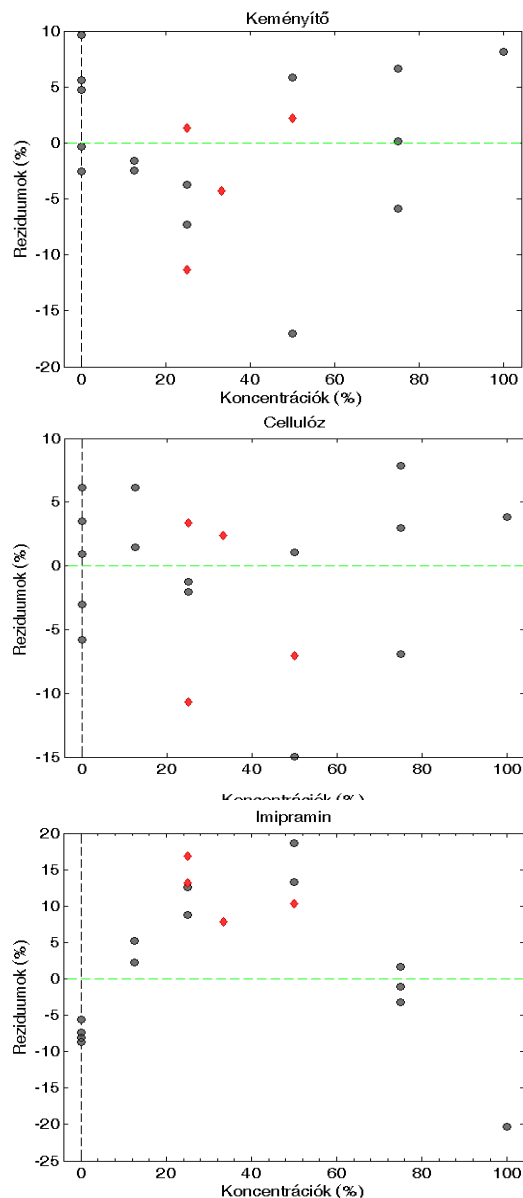
4.2.6. Nemlineáris modellek alkalmazása

A rendszerben a segédanyagok mennyiségi meghatározása során nem tapasztaltam jelentős eltérést a lineáris viselkedéstől, azonban egyes modelleknél a hatóanyag esetén erős nemlinearitás tapasztaltam (4.2.6. ábra). Ez a jelenség magyarázható azzal is, hogy a hatóanyag móltömege jelentősen eltér a makromolekulás segédanyagokétól, ami a tömegszázalékokkal történő kalibrálásban torzítást okozhat.

A nemlineáris viselkedést már a PLS modell is viszonylag jó kezelte (a reziduum ábrán tapasztalható nemlineáris trend csökkent), azonban felmerülhet az a kérdés, hogy a nemlineáris modellek használhatóak-e abban az esetben, ha a vizsgálni kívánt több komponens közül nem mindegyik esetben tapasztalható nemlineáris viselkedés.

A háromkomponensű rendszer kvantitatív meghatározására a polimorfok esetén sikerrel alkalmazott helyileg súlyozott regressziót illetve támogatóvektorok módszerét nem tudtam vizsgálni, mivel ezek a módszerek csak egy függő változót, azaz egy komponens

koncentrációértékeit tudják kezelni. Többkomponensű rendszer esetén arra van szükségünk, hogy a több függő változóval egyszerre, a köztük lévő kapcsolat minél pontosabb felderítésével tudjuk modellezni. Ilyen komplex rendszerek nemlineáris modellezésére alkalmas lehetőség lehet a mesterséges ideghálózatok módszere.



4.2.6. ábra: A CLS modellezés reziduumai

4.2.6.1. *Polinomiális PLS alkalmazása a háromkomponensű keverék mennyiségi meghatározására*

A PLS regresszió magasabb fokszámú összefüggéseit alkalmazó polinomiális PLS modellezést az imipraminos keverékre is tudtam alkalmazni. Az ilyen nemlineáris regresszióban az illesztésre használt polinom fokszámát a carvelilol rendszeren is bemutatott RMSE értékek minimalizálásával kerestem meg. A három látens változóval épített PLS modellek RMSE értékei

másodfokú függvény illesztésekor voltak minimálisak. A keményítő előrejelzési hibája nőtt a lineáris PLS modellhez képest, azonban a cellulóz és imipramin esetén a módszer hozzájárul a pontosabb meghatározáshoz, ez az imipramin RMSEP értékében 1%-os csökkenést jelent.

4.2.6.2. Mesterséges ideghálózatokkal történő modellezés

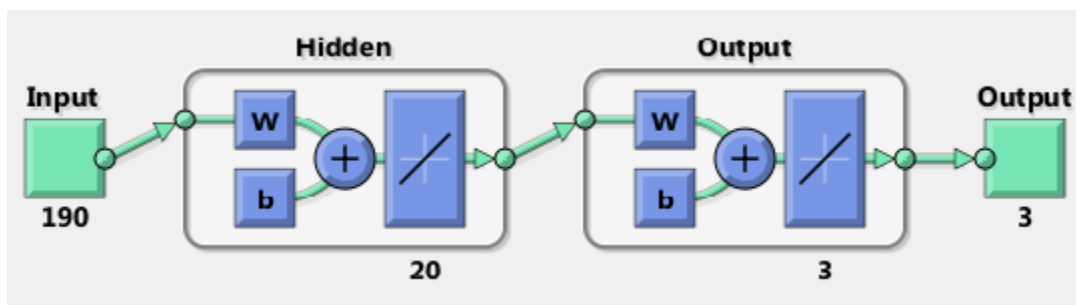
A 3.6.5.4. pontban bemutatott mesterséges ideghálózatok az adathalmazban lévő nemlinearitás kezelésére rugalmas lehetőséget nyújtanak, mivel nem szükséges meghatározni előre az illesztett polinom fokszámát, illetve több komponens esetén is használható. A módszer legfőbb hátránya, hogy a betanító halmaz elemeire tökéletesen illeszkedő modellt tud alkotni, azonban emiatt előrejelzésre ez nem használható. Így a fő szempont e módszer alkalmazása során a túlillesztés megakadályozása volt, melyre több lehetőség is rendelkezésre áll. A túlillesztés elkerülésének legegyszerűbb módja nagyszámú kalibráló pont használata, azonban erre sok esetben nincs lehetőség. Az általam használt adathalmaz is kicsinek bizonyult, ugyanis a 15 kalibráló pontot még az algoritmus tovább bontotta betanító, validáló és tesztelő mintákra is 50-25-25% -os arányban.

A túlillesztés problémáját úgy próbáltam megoldani, hogy nem a teljes spektrumokat használtam a mennyiségi meghatározáshoz, hanem csak 190 változót, melyet genetikusan algoritmus segítségével választottam ki. A kalibrálásra használt minták számát pedig úgy növeltem, hogy az eredeti kalibráló spektrumokat különböző nagyságú véletlen hibával terheltem, és a továbbiakban ezeket is egy-egy kalibráló spektrumként használtam. Ez az eljárás a kalibráció hibáját nyilvánvalóan növeli, illetve emiatt ez nem összehasonlítható a többi modellel. Így a modell jellemzésére a predikció hibáját illetve R^2 értékét használtam.

Olyan hálózatot teszteltem, mely egy rejtett réteget tartalmaz, a közbülső csomópontok számát tapasztalati úton határoztam meg. Az algoritmus egyéb beállításaitól függően ezt a számot 5 és 15 között találtam optimálisnak, ennél nagyobb érték esetén a modell előrejelző képessége nagyon leromlott. A csomópontok átalakító függvényeként pedig megvizsgáltam a lineáris, a tangens hiperbolikus és szigmoid típusú függvényeket is. Az utóbbi két esetben a függvények értékkészlete szerint (-1,1) illetve (0,1) közé skáláztam át a koncentrációkat. Azt tapasztaltam, hogy a lineáris átviteli függvényt használó hálózatok eredményezték a legkisebb előrejelzési hibát, mely magyarázható azzal, hogy a rendszerben viszonylag enyhe a nemlineáris

viselkedés. A betanításra több algoritmus kipróbálása után a Levenberg-Marquardt-féle visszafuttatásos algoritmust alkalmaztam. Az épített hálózat szerkezetét a 4.2.7. ábra szemlélteti.

A mesterséges ideghálózatok futása számos regularizációs/szabályozó paraméterrel befolyásolható, melyek hatással vannak a modell előrebecslő képességére is. Ezek részletes vizsgálatát további munkám során tervezem megvalósítani lehetőség szerint ezzel javítva az elérhető meghatározási pontosságot.



4.2.7. ábra: A mesterséges ideghálózat felépítése

A mesterséges ideghálók regressziós alkalmazása tehát nagy körütekintést igényel a nagyon gyakori és erős túlillesztés miatt, illetve a leginkább problémát jelentő tényező az, hogy sikeres alkalmazásához nagyszámú mérési adat szükséges. Ezért alkalmazása olyan technológiák esetén lehet sikeres, mely nagyszámú mérési adatot tud szolgáltatni a kalibráláshoz. Ugyanakkor a módszer alkalmazhatóságára utal az, hogy a kevés ponttal betanított hálózat előrebecslési hibája is a PLS modellel hasonló nagyságú értéket vett fel, a legjobb modell a keményítőt 3,1%, a cellulózt 4,9% az imipramint pedig 5,3% hibával becsülte.

4.2.7. A modellek összehasonlítása

A különböző módszerekkel nyert eredményeket a 4.2.3. táblázatban foglalom össze. A táblázatban megfigyelhetjük, hogy az egyes módszerek különbözőképpen hatnak a komponensek meghatározási hibáira. A táblázatban nem tüntettem fel a kapott regressziós egyenesek tengelymetszetét, mivel azoknak volt a legkisebb jelentősége a modellek jellemzésében.

A minden komponensre külön számolt, modelljóságot jellemző paraméterek nagy száma miatt a modellek összehasonlítása nehézkes. Azt, hogy melyik az „átlagosan legjobb modell” a változókiválasztási módszerek összehasonlítása során (4.2.5. pont) bemutatott rangsorolással döntöttem el. A rangsorolás eredményét a 4.2.3. táblázat utolsó sora tartalmazza. Megfigyelhető,

hogy az intervallum PLS-t és genetikus algoritmusokat alkalmazó PLS modellek adódtak a legjobbnak. Eszerint a sorba állítás szerint a PCR többször adott kisebb hibát a PLS-nél. A legpontatlanabb módszer a CLS és az egyváltozós regresszió. E rangsor azonban nem veszi figyelembe a túlillesztést, a modellek új mintákra való használhatóságára ebből nem következtethetünk.

A táblázat eredményei is mutatják, hogy míg az egyváltozós regresszió és a CLS modell mindhárom komponens megfelelő becslésére nem alkalmas, addig a PLS modellel, illetve változókiválasztással a segédanyagok és a hatóanyag is jelentősen pontosabban határozható meg. Az egyes modellek különböző mértékben javítják a komponensek meghatározásának pontosságát.

4.2.3. táblázat: A különböző módszerekkel nyert modellekkel nyert eredmények

		Egyváltozós	CLS	PCR	PLS	iPLS			PLS-GA	
						reverse, 5 változó/ablak	forward, 10 változó/ablak	reverse, 10 változó/ablak	10 változó/ablak	5 változó/ablak
RMSEC	kem.	9,36	6,82	6,03	6,09	4,87	4,20	3,99	5,21	5,62
	cell.	60,02	5,76	5,83	5,47	4,67	2,08	4,60	3,13	3,00
	imp.	11,68	10,04	6,33	5,67	3,46	4,03	3,71	4,90	3,95
RMSECV	kem.	-	8,50	7,77	8,10	6,95	5,94	6,96	7,08	7,69
	cell.	-	7,11	8,01	7,69	6,54	3,04	6,57	3,99	4,79
	imp.	-	12,60	8,71	8,37	5,61	6,20	7,62	7,28	7,39
RMSEP	kem.	7,38	6,17	6,48	6,53	5,28	6,50	5,45	6,72	5,37
	cell.	18,98	6,71	6,20	5,39	3,95	8,42	4,47	4,83	4,23
	imp.	15,10	12,51	4,28	3,90	4,37	2,43	4,03	4,73	2,50
R ² _cal	kem.	0,9275	0,9601	0,9674	0,9669	0,9788	0,9843	0,9857	0,9757	0,9718
	cell.	0,2871	0,9712	0,9696	0,9733	0,9805	0,9961	0,9811	0,9913	0,9919
	imp.	0,8908	0,9174	0,9641	0,9712	0,9893	0,9855	0,9877	0,9785	0,9861
R ² _cv	kem.	-	0,9406	0,9463	0,9417	0,9570	0,9685	0,9568	0,9551	0,9473
	cell.	-	0,9573	0,9431	0,9477	0,9617	0,9921	0,9614	0,9858	0,9795
	imp.	-	0,8753	0,9322	0,9374	0,9720	0,9657	0,9486	0,9525	0,9512
R ² _val	kem.	0,8178	0,8934	0,8690	0,8648	0,9319	0,9114	0,9379	0,9018	0,8764
	cell.	0,3646	0,6952	0,6898	0,7421	0,8601	0,5789	0,8256	0,8062	0,8434
	imp.	0,9509	0,9000	0,9000	0,9438	0,9602	0,9537	0,9464	0,9156	0,9654
rangsorolás eredménye		9.	8.	6.	7.	1.	2.	3.	5.	4.

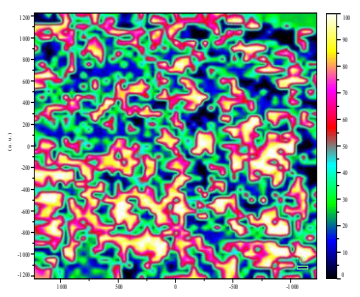
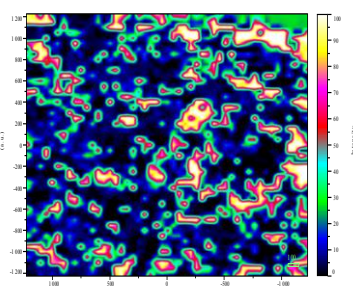
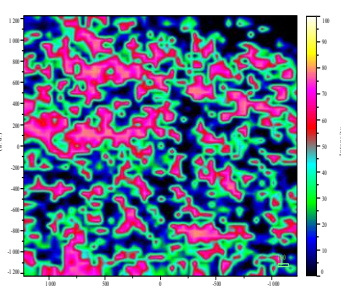
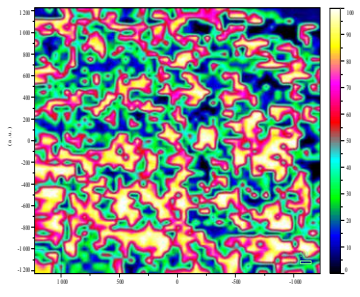
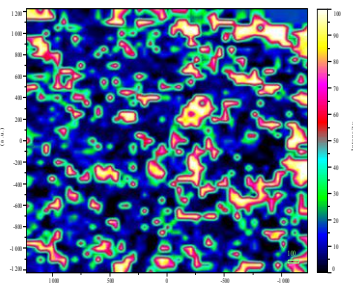
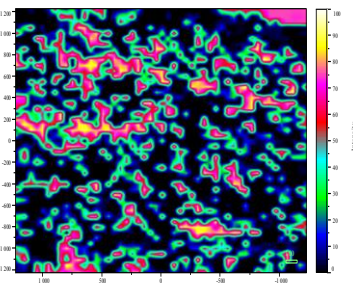
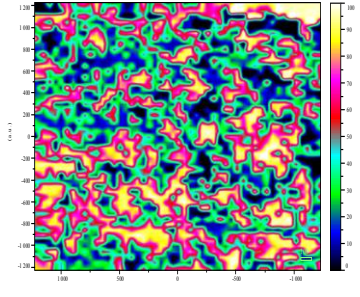
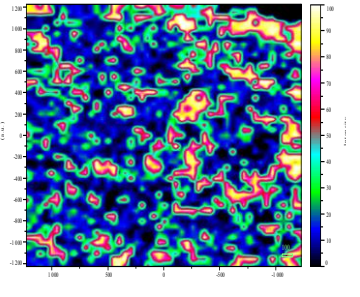
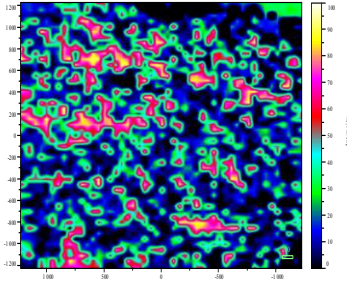
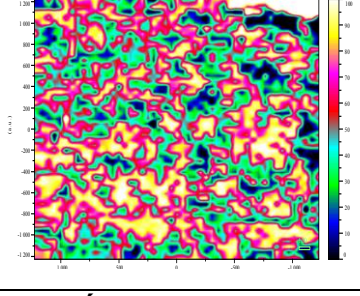
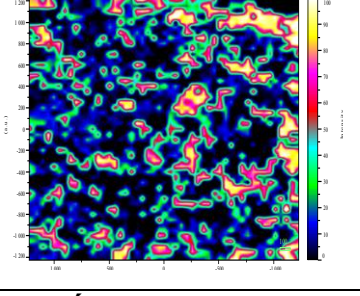
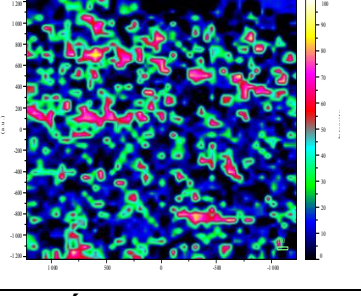
A mesterséges ideghálókkal épített modellek más kalibráló illetve validáló felosztást alkalmaztak, ezért ezeket nem kívántam összehasonlítani a táblázatban foglalt modellekkel, mert az félrevezető lehet. A táblázat ilyen megfontolásból az ANN modellek eredményeit nem tartalmazza.

A modellek összehasonlítására felmerülhet az SRD alkalmazásának lehetősége is, azonban ez a módszer a modelleknek azt képességét jellemzi a legjobban, hogy mennyire tudják a közeli koncentrációszinthez megkülönböztetni. Az ebben a részben vizsgált mintasorozat azonban nem tartalmaz ilyen közeli koncentrációszinthez, ezért az SRD alkalmazását jelen esetben elvettem.

4.2.7.1. A modellek alkalmazása új térképek mennyiségi elemzésére

A modellek előrejelző képességét az RMSEP értékeken kívül megvizsgáltam a modellek építéséhez nem használt térkép elemzése által is. A makropixel-elemzéshez már használt M9-es mintáról felvett 49×49 képpontból álló térképet különböző modellekkel értékeltem ki. A kapott koncentráció-térképeket illetve az abból számolt statisztikai jellemzőket a 4.2.8. ábrán láthatjuk.

Az ábrákon megfigyelhető, hogy a keményítő koncentráció-térképei közt viszonylag kis eltérések tapasztalhatók, míg a különbségek számottevőbbek a cellulóz és az imipramin esetén. Ahogy az RMSE értékek segítségével is megállapítottam, a CLS az imipramin mennyiségi meghatározására használható a legkevésbé, míg a cellulóz koncentrációját meglepő pontossággal, 1% alatti hibával adta vissza.

	Keményítő (50%)	Cellulóz (25%)	Imipramin (25%)
CLS			
	Átlag= 46,1 % Szórás= 34,53	Átlag= 25,70 % Szórás= 33,95	Átlag= 34,38 % Szórás= 29,34
PLS			
	Átlag= 49,31 % Szórás= 34,39	Átlag= 28,01 % Szórás= 33,44	Átlag= 22,68 % Szórás= 32,21
PLS-GA			
	Átlag= 46,05 % Szórás= 34,18	Átlag= 29,83 % Szórás= 32,40	Átlag= 24,12 % Szórás= 29,53
ANN			
	Átlag= 56,67 % Szórás= 33,73	Átlag= 24,61 % Szórás= 34,23	Átlag= 19,46 % Szórás= 25,71

4.2.8. ábra: A modellek által becsült koncentráció-térképek és statisztikai jellemzőik

A genetikus algoritmussal támogatott PLS modell a hatóanyag meghatározási pontosságát javította, míg a segédanyagok átlagkoncentrációját kevésbé jól becsülte, mint a PLS modell. Ez arra világít rá, hogy a változó kiválasztási módszerek valószínűleg az imipramin jellemző csúcsait tudják kiválasztani a hatóanyag spektrumán jelentkező sok és éles csúcsnak köszönhetően. Ezek kiválasztásával a kalibrációja is pontosabbá tehető. Ezzel szemben a segédanyagok mennyisége, széles sávjaik miatt, pontosabban leírható, ha az egész spektrumot felhasználjuk.

A cellulóz meghatározására a CLS modellezésen kívül a mesterséges ideghálózatok is jónak adódtak, mely módszer a többi komponenst a többi modellhez képest rosszul becsülte. Az ANN modell túlillesztettségét tehát nem sikerült tökéletesen kiküszöbölni, azonban annak ellenére, hogy kevés kalibráló pontot használtam a betanításra, a többi modellel összevethető pontosságot sikerült vele elérni. Ez is arra utal, hogy több mérési adat esetén a módszer alkalmas lehet pontos kvantitatív meghatározásra.

A hatóanyagot és segédanyagokat tartalmazó többkomponensű keverék rendszer mennyiségi meghatározását többváltozós regressziós módszerekkel végeztem el. Számos, irodalomból ismert, de Raman-térképeken kevésbé alkalmazott módszert használtam és vetettem össze. Az elvégzett kiértékelésekből megállapítható, hogy sikerült olyan modellt építeni, mellyel egyidejűleg becsülhető mindhárom komponens, 5% alá eső hibával. Ennek a hibaértéknek csökkentésére törekedtem nemlineáris és változó kiválasztási módszerek alkalmazásával, és azt állapítottam meg, hogy egyes komponensek becslései ezekkel pontosabbá tehetők, azonban mindhárom komponens meghatározási hibáját nem tudtam egyidejűleg jelentősen lecsökkenteni. Viszont ha ismert folyamatban szeretnénk pontos kvantitatív modellt építeni, lehetőségünk van olyan módszer választására, mely a technológia szempontjából legfontosabb komponenst nagyobb pontossággal becsli, miközben a többi komponens is meghatározható, valamivel nagyobb hibával.

5. Összefoglalás

A gyógyszeriparban napjainkban különösen nagy hangsúlyt helyeznek a roncsolásmentes, gyors és a gyártásba könnyen beiktatható analitikai eszközök fejlesztésére. A készítménygyártási folyamatokat a legújabb irányelvek szerint folyamatosan nyomon kell követni, melyre a Folyamatfelügyelő és Analizáló Technológiák, röviden PAT (Process Analytical Technology) keretein belül lehet megoldani. Ennek a kiépítésével mindig tisztában tudunk lenni a minőséget befolyásoló tényezők aktuális állapotával, és akár azonnali beavatkozásra is módunk nyílna, ha megfelelően ismerjük a folyamatainkat.

A Raman-térképezésnek számos olyan jellemzője van, amely alkalmassá teszi arra, hogy gyógyszeripari kontrollált folyamatokban lehessen használni. A módszer megfelelő használatához azonban elengedhetetlenek az olyan adatelemzési módszerek, melyek lehetővé teszik a kvantitatív és kvalitatív információk pontos kinyerését. Munkám során kétkomponensű polimorf keverék rendszer illetve egy hatóanyagot és két segédanyagot tartalmazó háromkomponensű rendszer mennyiségi meghatározása során vizsgáltam, hogy milyen adatelemzési technikákkal lehet megbízható eredményeket szolgáltató kvantitatív modelleket létrehozni. Mindkét összetétel típus elemzése fontos gyógyszeripari feladat, ugyanakkor kvantitatív elemzésük során különböző kérdések merülnek fel. Míg polimorfok esetén kis mennyiségű szennyezés felderítésére van szükség és a fő kihívás a spektrumok hasonlóságából ered, addig a hatóanyag-segédanyag rendszerben kettőnél több komponens egyidejű pontos meghatározása a fő cél.

A két rendszer elemzése során olyan, a gyógyszeriparban is elterjedten használt adatelemzési technikákat vizsgáltam meg, mint a legkisebb négyzetek módszere (CLS), a főkomponens-regresszió (PCR) és a részleges legkisebb négyzetek módszere (PLS) illetve összehasonlítottam ezeket a Lambert-Beer törvényen alapuló egyváltozós kiértékeléssel. Megállapítottam, hogy az adatelőkészítések megfelelő alkalmazása döntően befolyásolja a modell eredményességét, az alapvonal-korrekción és normálás mellett különösen hasznosnak az átlagra centrálást találtam. Azt tapasztaltam, hogy efféle kemometriai módszerek alkalmazásával a nem, vagy csak nagy hibával meghatározható komponensek pontos mennyiségi elemzése is lehetővé válik.

A gyakran tapasztalható nemlineáris viselkedés kezelésére az említett módszerek nem alkalmasak. A kvantitatív elemzésre ilyen esetekben nemlineáris módszerek alkalmazása javasolt.

Munkám során ezek közül a polinomiális PLS-t, a támogatóvektorok módszerét (SVM), a helyileg súlyozott regressziót (LWR) valamint a mesterséges ideghálózatokat (ANN) egyaránt alkalmaztam. Az eredmények igazolták, hogy nemlinearitás jelenléte esetén célszerű ezeknek a módszereknek a használata.

Különös figyelmet fordítottam a változókiválasztási módszerek tesztelésére, melyek alkalmazására Raman-térképezés területén nem található példa. Mind az intervallum-PLS, mind a genetikusan algoritmus alkalmas módszernek bizonyult a meghatározás szempontjából legfontosabb hullámszám-csatornák kiválasztására, valamint a nemlinearitást okozó sávok eltávolítására. A változókiválasztások alkalmazása különösen javasolt lehet kontrollált technológiák esetén, mivel amellyel, hogy pontosabb meghatározást tesznek lehetővé, rámutatnak arra is, hogy melyik az a spektrális tartomány, amelyet érdemes felvenni, s ezáltal akár a térképezés ideje is csökkenthető.

A felépített modelleket új, a modellépítéshez nem használt Raman-térképek mennyiségi elemzésére használtam. Meggyőződtem arról, hogy az alkalmazott módszerekkel pontos előrejelző képességű modelleket tudtam építeni. Ezek a térképek lehetőséget adtak homogenitásvizsgálat elvégzésére is, mely PAT körülmények között szintén fontos kérdésekre világíthat rá, például az optimális térképméret megválasztásában is segítséget nyújthat. A Raman-térképezés és pontos mennyiségi modellek kombinációját azért is lehet célszerű PAT-ban alkalmazni, mert akár olyan kis szennyeződéseket is ki tud mutatni a részletes térbeli felbontásnak köszönhetően, amely lokális kimutatási határ felett van, de globálisan csak néhány tized százalékot tesz ki.

A kvantitatív modellek építése során nyert eredmények rámutattak arra, hogy a megfelelő kemometria módszerek alkalmazásával lehetőség nyílik olyan komponensek pontos meghatározására is, melyek meghatározása ilyen pontossággal mindeztáig nem volt megoldott. A dolgozatomban bemutatott tapasztalatok ezek alapján hozzájárulhatnak a Raman-térképezéssel végzett kvantitatív analízis továbbfejlesztéséhez, s annak integrálásához a Folyamatfelügyelő és Analizáló Technológiába.

6. Köszönetnyilvánítás

Köszönetet szeretnék mondani mindazoknak, akik munkám során segítséget nyújtottak. Elsősorban szeretném megköszönni konzulensemnek, Farkas Attilának lelkiismeretes munkáját és türelmes irányítását. Köszönetet mondok továbbá Dr. Madarász Jánosnak a röntgen-diffrakciós felvételekért, valamint a tanszéki kutatócsoport minden tagjának a munkám során nyújtott segítségért és tanácsokért. Végül, de nem utolsósorban megköszönöm a támogatást és biztatást a mellettem álló családomnak és barátaimnak.

7. Irodalomjegyzék

- [1] L. Brillouin *Ann. Phys.* 88, 17 **(1922)**
- [2] A. Smekal *Naturwissenschaften* 11, 873 **(1923)**
- [3] C.V. Raman, K.S. Krishnan: A New Type of Secondary Radiation, *Nature* 121, 501-2 **(1928)**
- [4] Placzek, G. *Handbuch der Radiologie*, Vol. 6 Pt.2; Akad. Verlagsgesellschaft, Leipzig 1934 p 205
- [5] A. Schawlow, C. H. Townes *Phys. Rev.* 122, 1940 **(1958)**
- [6] T. H. Maiman *Nature* 187, 493 **(1960)**
- [7] G. J. Rosasco, E. S. Etz, W. A. Cassatt *IVth Int. Conf. Raman Spectrosc.* Brunswick USA **(1974)**
- [8] M. Delhaye, P. Dhamelincourt *IVth Int. Conf. Raman Spectrosc.* Brunswick USA **(1974)**
- [9] A.A. Gowen, C.P. O'Donnell, P.J. Cullen, S.E.J. Bell, Recent applications of Chemical Imaging to pharmaceutical process monitoring and quality control, *European Journal of Pharmaceutics and Biopharmaceutics*, 69 **(2008)** 10-22.
- [10] C. Gendrin, Y. Roggo, C. Collet, Pharmaceutical applications of vibrational chemical imaging and chemometrics: A review, *Journal of Pharmaceutical and Biomedical Analysis*, 48 **(2008)** 533-553.
- [11] K.C. Gordon, C.M. McGoverin, Raman mapping of pharmaceuticals, *International Journal of Pharmaceutics*, 417 **(2011)** 151-162.
- [12] J.M. Amigo, J. Cruz, M. Bautista, S. MasPOCH, J. Coello, M. Blanco, Study of pharmaceutical samples by NIR chemical-image and multivariate analysis, *TrAC Trends in Analytical Chemistry*, 27 **(2008)** 696-713.
- [13] P.Y. Sacré, C. De Bleye, P.F. Chavez, L. Netchacovitch, P. Hubert, E. Ziemons, Data processing of vibrational chemical imaging for pharmaceutical applications, *Journal of Pharmaceutical and Biomedical Analysis*.
- [14] Andrew, J.J., Hancewicz, T.M., Rapid analysis of Raman image data using two- way multivariate curve resolution. *Appl. Spectrosc.* 52 **(1998)**, 797–807.
- [15] Sasic, S.; Clark, D. A.; Mitchell, J. C.; Snowden, M. J. Univariate versus multivariate Raman imaging- A simulation with an example from pharmaceutical practice, *Analyst* 129 **(2004)** 1001-1007.

- [16] N. Bergner, T. Bocklitz, B.F.M. Romeike, R. Reichart, R. Kalff, C. Krafft, J. Popp, Identification of primary tumors of brain metastases by Raman imaging and support vector machines, *Chemometrics and Intelligent Laboratory Systems*, 117 **(2012)** 224-232.
- [17] K. Klein, Alexander M. Gigler, T. Aschenbrenner, R. Monetti, W. Bunk, F. Jamitzky, G. Morfill, Robert W. Stark, J. Schlegel, Label-Free Live-Cell Imaging with Confocal Raman Microscopy, *Biophysical Journal*, 102 **(2012)** 360-368.
- [18] D. Wu, D.-W. Sun, Advanced applications of hyperspectral imaging technology for food quality and safety analysis and assessment: A review — Part I: Fundamentals, *Innovative Food Science & Emerging Technologies*, 19 **(2013)** 1-14.
- [19] D. Wu, D.-W. Sun, Advanced applications of hyperspectral imaging technology for food quality and safety analysis and assessment: A review — Part II: Applications, *Innovative Food Science & Emerging Technologies*, 19 **(2013)** 15-28.
- [20] S. Serranti, A. Gargiulo, G. Bonifazi, Characterization of post-consumer polyolefin wastes by hyperspectral imaging for quality control in recycling processes, *Waste Management*, 31 **(2011)** 2217-2227.
- [21] G.J. Edelman, E. Gaston, T.G. van Leeuwen, P.J. Cullen, M.C.G. Aalders, Hyperspectral imaging for non-contact analysis of forensic traces, *Forensic Science International*, 223 **(2012)** 28-39.
- [22] FDA, Guidance for Industry: PAT – A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance, **2004**.
<http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm070305.pdf>.
- [23] T. De Beer, A. Burggraef, M. Fonteyne, L. Saerens, J.P. Remon, C. Vervaet, Near infrared and Raman spectroscopy for the in-process monitoring of pharmaceutical production processes, *International Journal of Pharmaceutics*, 417 **(2011)** 32-47.
- [24] A.L. Pomerantsev, O.Y. Rodionova, Process analytical technology: a critical view of the chemometricians, *Journal of Chemometrics*, 26 **(2012)** 299-310.
- [25] P. Geladi: Chemometrics in spectroscopy. Part 1. Classical chemometrics, *Spectrochim. Acta Pt B*, 58 **(2003)**, 767-782
- [26] T. Rajalahti, O.M. Kvalheim, Multivariate data analysis in pharmaceuticals: A tutorial review, *International Journal of Pharmaceutics*, 417 **(2011)** 280-290.

- [27] G. Stahly, Diversity in single- and multiple-component crystals. The search for and prevalence of polymorphs and cocrystals, *Cryst. Growth Des.* 7 **(2007)** 1007–1026.
- [28] N. Chieng, T. Rades, J. Aaltonen, An overview of recent studies on the analysis of pharmaceutical polymorphs, *Journal of Pharmaceutical and Biomedical Analysis*, 55 **(2011)** 618-644.
- [29] S.N. Campbell Roberts, A.C. Williams, I.M. Grimsey, S.W. Booth, Quantitative analysis of mannitol polymorphs. FT-Raman spectroscopy, *Journal of Pharmaceutical and Biomedical Analysis*, 28 **(2002)** 1135-1147.
- [30] Z. Németh, G.C. Kis, G. Pokol, Á. Demeter, Quantitative determination of famotidine polymorphs: X-ray powder diffractometric and Raman spectrometric study, *Journal of Pharmaceutical and Biomedical Analysis*, 49 **(2009)** 338-346.
- [31] D.M. Croker, M.C. Hennigan, A. Maher, Y. Hu, A.G. Ryder, B.K. Hodnett, A comparative study of the use of powder X-ray diffraction, Raman and near infrared spectroscopy for quantification of binary polymorphic mixtures of piracetam, *Journal of Pharmaceutical and Biomedical Analysis*, 63 **(2012)** 80-86.
- [32] K. Kachrimanis, D.E. Braun, U.J. Griesser, Quantitative analysis of paracetamol polymorphs in powder mixtures by FT-Raman spectroscopy and PLS regression, *Journal of Pharmaceutical and Biomedical Analysis*, 43 **(2007)** 407-412.
- [33] N. Chieng, S. Rehder, D. Saville, T. Rades, J. Aaltonen, Quantitative solid-state analysis of three solid forms of ranitidine hydrochloride in ternary mixtures using Raman spectroscopy and X-ray powder diffraction, *Journal of Pharmaceutical and Biomedical Analysis*, 49 **(2009)** 18-25.
- [34] C.M. McGoverin, L.C.H. Ho, J.A. Zeitler, C.J. Strachan, K.C. Gordon, T. Rades, Quantification of binary polymorphic mixtures of ranitidine hydrochloride using NIR spectroscopy, *Vibrational Spectroscopy*, 41 **(2006)** 225-231.
- [35] C.M. McGoverin, M.D. Hargreaves, P. Matousek, K.C. Gordon, Pharmaceutical polymorphs quantified with transmission Raman spectroscopy, *Journal of Raman Spectroscopy*, 43 **(2012)** 280-285.
- [36] Li, Y.; Chow, P. S.; Tan, R. B. H., Quantification of polymorphic impurity in an enantiotropic polymorph system using differential scanning calorimetry, X-ray powder diffraction and Raman spectroscopy, *Int. J. Pharm.* 415 **(2011)** 110-118.

- [37] S.A. Schönrichler, L.K.H. Bittner, A.K.H. Weiss, U.J. Griesser, J.D. Pallua, C.W. Huck, Comparison of NIR chemical imaging with conventional NIR, Raman and ATR-IR spectroscopy for quantification of furosemide crystal polymorphs in ternary powder mixtures, *European Journal of Pharmaceutics and Biopharmaceutics*, 84 **(2013)** 616-625.
- [38] Hu, Y.; Erxleben, A.; Ryder, A. G.; McArdle, P., Quantitative analysis of sulfathiazole polymorphs in ternary mixtures by attenuated total reflectance infrared, near-infrared and Raman spectroscopy *J. Pharm. Biomed. Anal.* **(2010)**, 53, 412-420.
- [39] Hellsten, S.; Qu, H.; Heikkilä, T.; Kohonen, J.; Reinikainen, S.-P.; Louhi-Kultanen, M., 'Raman spectroscopic imaging of indomethacin loaded in porous silica *CrystEngComm* **2012**, 14, 1582-1587.
- [40] Nielsen, L. H.; Keller, S. S.; Gordon, K. C.; Boisen, A.; Rades, T.; Müllertz, A., Spatial confinement can lead to increased stability of amorphous indomethacin, *Eur. J. Pharm. Biopharm.* **(2012)**, 81, 418-425.
- [41] A. Heinz, M. Savolainen, T. Rades, C.J. Strachan, Quantifying ternary mixtures of different solid-state forms of indomethacin by Raman and near-infrared spectroscopy, *European Journal of Pharmaceutical Sciences*, 32 **(2007)** 182-192.
- [42] W. Fortunato de Carvalho Rocha, G.P. Sabin, P.H. Marçó, R.J. Poppi, Quantitative analysis of piroxicam polymorphs pharmaceutical mixtures by hyperspectral imaging and chemometrics, *Chemometrics and Intelligent Laboratory Systems*, 106 **(2011)** 198-204.
- [43] M.C. Hennigan, A.G. Ryder, Quantitative polymorph contaminant analysis in tablets using Raman and near infra-red spectroscopies, *Journal of Pharmaceutical and Biomedical Analysis*, 72 **(2013)** 163-171.
- [44] S. Piqueras, L. Duponchel, R. Tauler, A. de Juan, Monitoring polymorphic transformations by using in situ Raman hyperspectral imaging and image multiset analysis, *Analytica Chimica Acta*, 819 **(2014)** 15-25.
- [45] F.L.F. Soares, R.L. Carneiro, Evaluation of analytical tools and multivariate methods for quantification of co-former crystals in ibuprofen-nicotinamide co-crystals, *Journal of Pharmaceutical and Biomedical Analysis*, 89 **(2014)** 166-175.
- [46] T. Okumura, M. Nakazono, M. Otsuka, K. Takayama, An accurate quantitative analysis of polymorphs based on artificial neural networks, *Colloids and Surfaces B: Biointerfaces*, 49 **(2006)** 153-157.

- [47] C. Ravn, E. Skibsted, R. Bro, Near-infrared chemical imaging (NIR-CI) on pharmaceutical solid dosage forms—Comparing common calibration approaches, *Journal of Pharmaceutical and Biomedical Analysis*, 48 **(2008)** 554-561.
- [48] A.D. Karande, P.W.S. Heng, C.V. Liew, In-line quantification of micronized drug and excipients in tablets by near infrared (NIR) spectroscopy: Real time monitoring of tableting process, *International Journal of Pharmaceutics*, 396 **(2010)** 63-74.
- [49] J.G. Rosas, M. Blanco, A criterion for assessing homogeneity distribution in hyperspectral images. Part 1: Homogeneity index bases and blending processes, *Journal of Pharmaceutical and Biomedical Analysis*, 70 **(2012)** 680-690.
- [50] J.G. Rosas, M. Blanco, A criterion for assessing homogeneity distribution in hyperspectral images. Part 2: Application of homogeneity indices to solid pharmaceutical dosage forms, *Journal of Pharmaceutical and Biomedical Analysis*, 70 **(2012)** 691-699.
- [51] C. Gendrin, Y. Roggo, C. Collet, Content uniformity of pharmaceutical solid dosage forms by near infrared hyperspectral imaging: A feasibility study, *Talanta*, 73 **(2007)** 733-741.
- [52] G.P. Sabin, W.F. de Carvalho Rocha, R.J. Poppi, Study of the similarity between distribution maps of concentration in near-infrared spectroscopy chemical imaging obtained by different multivariate calibration approaches, *Microchemical Journal*, 99 **(2011)** 542-547.
- [53] N. Scoutaris, K. Vithani, I. Slipper, B. Chowdhry, D. Douroumis, SEM/EDX and confocal Raman microscopy as complementary tools for the characterization of pharmaceutical tablets, *International Journal of Pharmaceutics*, 470 **(2014)** 88-98.
- [54] Y. Sulub, B. Wabuyeale, P. Gargiulo, J. Pazdan, J. Cheney, J. Berry, A. Gupta, R. Shah, H. Wu, M. Khan, Real-time on-line blend uniformity monitoring using near-infrared reflectance spectrometry: A noninvasive off-line calibration approach, *Journal of Pharmaceutical and Biomedical Analysis*, 49 **(2009)** 48-54.
- [55] P.R. Wahl, G. Fruhmann, S. Sacher, G. Straka, S. Sowinski, J.G. Khinast, PAT for tableting: Inline monitoring of API and excipients via NIR spectroscopy, *European Journal of Pharmaceutics and Biopharmaceutics*, 87 **(2014)** 271-278.
- [56] L.B. Lyndgaard, F. van den Berg, A. de Juan, Quantification of paracetamol through tablet blister packages by Raman spectroscopy and multivariate curve resolution-alternating least squares, *Chemometrics and Intelligent Laboratory Systems*, 125 **(2013)** 58-66.

- [57] A. Palou, J. Cruz, M. Blanco, J. Tomàs, J. de los Ríos, M. Alcalà, Determination of drug, excipients and coating distribution in pharmaceutical tablets using NIR-CI, *Journal of Pharmaceutical Analysis*, 2 **(2012)** 90-97.
- [58] S. Grangeiro Junior, L. de Moura França, M.F. Pimentel, M.M. Albuquerque, D.P. de Santana, A.K.M. Santana, J.A.L. Souza, S.S. Simões, A process analytical technology approach for the production of fixed-dose combination tablets of zidovudine and lamivudine using near infrared spectroscopy and chemical images, *Microchemical Journal*.
- [59] H. Nakagawa, M. Kano, S. Hasebe, T. Miyano, T. Watanabe, N. Wakiyama, Verification of model development technique for NIR-based real-time monitoring of ingredient concentration during blending, *International Journal of Pharmaceutics*, 471 **(2014)** 264-275.
- [60] S. Kim, M. Kano, H. Nakagawa, S. Hasebe, Estimation of active pharmaceutical ingredients content using locally weighted partial least squares and statistical wavelength selection, *International Journal of Pharmaceutics*, 421 **(2011)** 269-274.
- [61] Z. Xiaobo, Z. Jiewen, M.J.W. Povey, M. Holmes, M. Hanpin, Variables selection methods in near-infrared spectroscopy, *Analytica Chimica Acta*, 667 **(2010)** 14-32.
- [62] T. Mehmood, K.H. Liland, L. Snipen, S. Sæbø, A review of variable selection methods in Partial Least Squares Regression, *Chemometrics and Intelligent Laboratory Systems*, 118 **(2012)** 62-69.
- [63] R. Leardi, A. Lupiáñez González, Genetic algorithms applied to feature selection in PLS regression: how and when to use them, *Chemometrics and Intelligent Laboratory Systems*, 41 **(1998)** 195-207.
- [64] R. Leardi, M.B. Seasholtz, R.J. Pell, Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data, *Analytica Chimica Acta*, 461 **(2002)** 189-200.
- [65] L. Srathaphut, N. Ruangwises, Genetic Algorithms-Based Approach for Wavelength Selection in Spectrophotometric Determination of Vitamin B12 in Pharmaceutical Tablets by Partial Least-Squares, *Procedia Engineering*, 32 **(2012)** 225-231.
- [66] L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck and S.B. Engelsen, Interval Partial Least Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy, *Applied Spectroscopy*, 54 **(2000)** 413-419
- [67] Leardi R, Nørgaard L. Sequential application of backward interval PLS and Genetic Algorithms for the selection of relevant spectral regions, *Journal of Chemometrics*, 18(11):486-497 **(2004)**

- [68] Y.-H. Yun, D.-S. Cao, M.-L. Tan, J. Yan, D.-B. Ren, Q.-S. Xu, L. Yu, Y.-Z. Liang, A simple idea on applying large regression coefficient to improve the genetic algorithm-PLS for variable selection in multivariate calibration, *Chemometrics and Intelligent Laboratory Systems*, 130 **(2014)** 76-83.
- [69] C. Ferrari, G. Foca, A. Ulrici, Handling large datasets of hyperspectral images: Reducing data size without loss of useful information, *Analytica Chimica Acta*, 802 **(2013)** 29-39.
- [70] C. Esquerre, A. Gowen, G. Downey, C. O'Donnell, Wavelength selection for development of a near infrared imaging system for early detection of bruise damage in mushrooms (*Agaricus bisporus*) *Journal of Near Infrared Spectroscopy*, 20 **(2012)** 537-546.
- [71] W.S. Cleveland: Robust Locally Weighted Regression and Smoothing Scatterplots, *J. Am. Stat. Assoc* 74 **(1979)** 829-836.
- [72] D. Pérez-Marín, A. Garrido-Varo, J.E. Guerrero, Non-linear regression methods in NIRS quantitative analysis, *Talanta*, 72 **(2007)** 28-42.
- [73] W. Ni, L. Nørgaard, M. Mørup, Non-linear calibration models for near infrared spectroscopy, *Analytica Chimica Acta*, 813 **(2014)** 1-14.
- [74] M.R. Khoshayand, H. Abdollahi, M. Shariatpanahi, A. Saadatfard, A. Mohammadi, Simultaneous spectrophotometric determination of paracetamol, ibuprofen and caffeine in pharmaceuticals by chemometric methods, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 70 **(2008)** 491-499.
- [75] M.R. Sohrabi, P. Abdolmaleki, M. Dehroudi, Simultaneous quantitative determination of Amlodipine and Atorvastatin in tablets using artificial neural networks, *Mathematical and Computer Modelling*, 58 **(2013)** 1588-1594.
- [76] H.W. Darwish, S.A. Hassan, M.Y. Salem, B.A. El-Zeany, Different approaches in Partial Least Squares and Artificial Neural Network models applied for the analysis of a ternary mixture of Amlodipine, Valsartan and Hydrochlorothiazide, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 122 **(2014)** 744-750.
- [77] M.M. Elkhoudary, R.A. Abdel Salam, G.M. Hadad, Comparative artificial neural network and partial least squares models for analysis of Metronidazole, Diloxanide, Spiramycin and Clotrimazole in pharmaceutical preparations, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 130 **(2014)** 222-229.

- [78] S. Mazurek, R. Szostak, Quantification of atorvastatin calcium in tablets by FT-Raman spectroscopy, *Journal of Pharmaceutical and Biomedical Analysis*, 49 **(2009)** 168-172.
- [79] B. Özbalci, İ.H. Boyacı, A. Topcu, C. Kadılar, U. Tamer, Rapid analysis of sugars in honey by processing Raman spectrum using chemometric methods and artificial neural networks, *Food Chemistry*, 136 1444-1452.
- [80] D.E. Braun, S.G. Maas, N. Zencirci, C. Langes, N.A. Urbanetz, U.J. Griesser, Simultaneous quantitative analysis of ternary mixtures of d-mannitol polymorphs by FT-Raman spectroscopy and multivariate calibration models, *International Journal of Pharmaceutics*, 385 **(2010)** 29-36.
- [81] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, USA, 1995.
- [82] V. Vapnik, *Statistical Learning Theory*, Wiley, New York, USA, **1998**.
- [83] U. Thissen, M. Pepers, B. Üstün, W.J. Melssen, L.M.C. Buydens, Comparing support vector machines to PLS for spectral regression applications, *Chemometrics and Intelligent Laboratory Systems*, 73 **(2004)** 169-179.
- [84] I.A. Naguib, E.A. Abdelaleem, M.E. Draz, H.E. Zaazaa, Linear support vector regression and partial least squares chemometric models for determination of Hydrochlorothiazide and Benazepril hydrochloride in presence of related impurities: A comparative study, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 130 **(2014)** 350-356.
- [85] K. Héberger, Sum of ranking differences compares methods or models fairly, *TrAC, Trends Anal. Chem.* 29 **(2010)** 101-109.
- [86] P. Eilers, A perfect smoother, *Anal. Chem.* 75 **(2003)** 3631-3636
- [87] A. Smola, B. Schölkopf, A tutorial on support vector regression, *Statistics and Computing*, 14 (2004) 199-222.