PLANT RESEARCH INTERNATIONAL

Analyse spectrale gegevens valplekken

M.M.W.B. Hendriks

Plant Research International B.V., Wageningen
november 2002

Nota 211

WAGENINGEN UR

Plant Research International B.V.

Adres	:	Droevendaalsesteeg 1, Wageningen
	:	Postbus 16, 6700 AA Wageningen
Tel.	:	0317 - 47 70 00
Fax	:	0317 - 41 80 94
E-mail	:	post@plant.wag-ur.nl
Internet	:	http://www.plant.wageningen-ur.nl

Inhoudsopgave

Samenvatting

De mogelijkheden van cropscan metingen voor classificatie van valplekken in gewas zijn onderzocht. Daartoe waren gegevens beschikbaar van cropscan metingen verricht aan valplekken in aardappelvelden. Elk van de valplekken was bovendien gekarakteriseerd als het resultaat van één van drie oorzaken: aardappelmoeheid, levende nematoden of structuurbederf.

Met behulp van twee verschillende multivariate classificatie technieken (PLS-DA en SIMCA) is gezocht naar relaties tussen de spectrale gegevens van de cropscan metingen en de oorzaak van de valplek. Zowel SIMCA als PLS-DA blijken niet discriminerend genoeg te zijn om een onderscheid te kunnen maken tussen de drie gedefinieerde valplek categorieën op basis van de spectrale waarnemingen. Dit wordt mede veroorzaakt door de grote verscheidenheid aan oorzaken die in de derde categorie, 'structuurbederf', vallen. De resultaten voor het onderscheiden van valplekken veroorzaakt door aardappelmoeheid ten opzichte van de beide andere categorieën is beter (LOO percentages correct geclassificeerde objecten > 80 %).

Er worden aanbevelingen gedaan voor verder onderzoek.

Summary

The potential of using cropscan measurements for classification of infestation foci in crops was explored. For that purpose data was available of cropscan measurement of potato crops. In addition each of the infestation foci was characterised as the result of one of three causes: potato sickness, living nematodes, or structure deterioration.

Relations between the spectral cropscan data and the cause of the infestation focus were sought with the help of two different multivariate classification techniques (PLS-DA and SIMCA). It turned out that the relations found with SIMCA as well as PLS-DA do not discriminate enough between the three different categories. This is also caused by the high diversity of causes, which are labelled as soil structure deterioration. Discriminating the first category infestation foci, potato sickness, from the other two together shows better results (LOO percentages correctly classified objects > 80%).

Recommendations are given for further research.

Introductie en beschrijving data

Door de Plantenziektekundige Dienst (PD) zijn velden met zgn. valplekken geïdentificeerd in aardappelvelden. Hierna zijn op deze velden met behulp van zgn. cropscan apparatuur spectrale reflectie metingen verricht. Een cropscan bevat een aantal sensoren waarbij voor bepaalde golflengten (ranges) de spectrale reflectie van het gewas gemeten wordt. De cropscan registreert 16 reflectiesignalen van steeds 1 m^2 oppervlak.

Op elk van de geïdentificeerde velden zijn cropscan metingen gedaan aan een braak liggend stuk land, een plek met gezond gewas, aan het centrum van de valplek en aan de rand van de valplek. Steeds werden 5 verschillend aaneengesloten stukken van 1 m^2 bemonsterd door 5 keer te scannen. Daarnaast zijn de verschillende valplekken ook nog gekarakteriseerd, als zijnde het gevolg van:

1. Aardappelmoeheid,
2. Levende nematoden (zitten m.n. aan de randen van de haard).
3. Structuurbederf (bodem / grondeigenschappen), in de totale dataset zitten in totaal 7 verschillende soorten structuurbederfplekken, als volgt gecodeerd voor gebruik in figuren:
    3. mestplek
    4. gemengwoeld
    5. waterschade
    6. structuurbederf, oorzaak onbekend
    7. onvoldoende bemest
    8. brandplek
    9. spuitspoor

In totaal zitten er in de PD database gegevens van 61 verschillende velden, waarop een groot aantal verschillende rassen stonden, en bovendien verschilden de ontwikkelingsstadia van het gewas ook nog per veld. Hierbij dient ook opgemerkt te worden dat elk van de rassen een ander resistentie en tolerantie heeft tegen aaltjes.

Analyses

De spectrale gegevens van elke plek zijn in vijfvoud verzameld, daardoor zijn er cropscan gegevens van vijf aaneengesloten plekken van 1 m^2. Deze gegevens zijn niet onafhankelijk, waarom besloten is de spectrale gegevens van de vijf plekken te middelen. In elk veld zijn zowel een stuk braak land gescand, een stuk gezond gewas, een plek in het midden van de valplek en aan de rand van de valplek. Voor elk van deze verschillende stukken van het veld zijn de spectra afzonderlijk gemiddeld, echter indien er op één veld meerdere valplekken gescand zijn, werden deze valplekken ook gemiddeld, en worden als een apart veld behandeld voor de analyses. Dit levert in totaal 33 velden / valplekken op in 2000 en 31 velden in 2001, onderverdeeld in 27 categorie 1 valplekken, 13 categorie 2 valplekken en 24 categorie 3 valplekken (alle vormen van structuurbederf).
In de Appendix zijn de ruwe data weergegeven in Figuur 1a voor 2000, in elke subplot zijn de ruwe gegevens voor 1 veld opgenomen (5 spectrale patronen voor elk van de plekken, gezond, braak, centrum valplek en rand valplek). In Figuur 1b zijn dezelfde gegevens uitgezet, maar nu voor 2001. In Figuur 1c zijn de gemiddelden van de vijf spectrale patronen weergegeven, echter nu uitgesplitst naar type plek en gecodeerd voor soort valplek (zie legenda bij de figuren voor codering van de verschillende kleuren).

Voor het analyseren van de spectrale gegevens moet er rekening gehouden worden met de invloed van de verschillende rassen en groeistadia op de spectrale reflectie. Bovendien kunnen de valplekken zo slecht zijn dat de kale grond een rol gaat spelen in de spectrale reflectie metingen. Meerdere manieren van corrigeren zijn bestudeerd, de correctie beschreven in deze nota bestaat uit het maken van een verschilspectrum van het spectrum van gezond gewas en het spectrum van de valplek (centrum of rand).

Voor de analyse van de spectrale gegevens is in eerste instantie gekozen voor twee verschillende methoden, PLS-DA (partial least squares – discriminant analyse), en de SIMCA methode. Beide methoden worden veelvuldig gebruikt voor het analyseren van multivariate gegevens wanneer het doel is klassen te onderscheiden.

PLS-DA is een PLS versie van discriminant analyse, waarbij er gebruik gemaakt wordt van het voordeel van PLS, nl. dat rekening gehouden wordt met colineariteit van verschillende variabelen. Voor een gedetailleerde beschrijving van PLS-DA zie o.a. Sjöström *et al.* [1]. Leave-one-out (LOO) kruisvalidatie werd gebruikt om het optimale aantal PLS componenten te bepalen. Als evaluatie criterium werd zowel de gesommeerde kwadratische voorspelfout (SEP), als het percentage correct geclassificeerde velden gebruikt. Voor de classificatie werd een indicator matrix gebruikt, waarin voor elke klasse een kolom werd opgenomen, waarbij indien een valplek in een bepaalde klasse viel dit gecodeerd werd met een 1, indien een valplek niet in betreffende klasse viel werd deze in betreffende kolom aangeduid met -1. Bij de voorspelling van de klasse van een specifieke valplek werd gezocht naar het maximum van de voorspelde waarden in de verschillende kolommen, en werd de valplek daarin geclassificeerd. Op deze manier kan in ieder geval elke voorspelling aan een bepaalde klasse gekoppeld worden.

SIMCA (soft independent method of class analogy) is een 'supervised' patroonherkennings methode. Een SIMCA model is een verzameling van PCA modellen, voor elke klasse één. Op basis van afstandsmaten van nieuwe objecten wordt bepaald of zo'n object behorende tot elk van de verschillende klassen gerekend kan worden. Dit betekent wel dat een nieuw object ook tot verschillende klassen gerekend kan worden, of tot geen van de klassen. Voor een uitgebreidere beschrijving van SIMCA zie de literatuur [2,3]. Voor het selecteren van het aantal principale componenten voor elke klasse wordt ook weer gebruik gemaakt van kruisvalidatievoorspellingen.

Alle analyses zijn uitgevoerd met behulp van het software pakket Matlab [4], gebruik makend van de PLS Toolbox [5].

Resultaten en discussie

Exploratief

In de Appendix, Figuren 1a en 1b zijn de ruwe data weergegeven, uitgesplitst naar veld. De figuren laten zien dat over het algemeen de centra van de valplekken en de randen van de valplekken een reductie geven van het spectrale patroon van de gezonde plekken op betreffend veld. In een aantal gevallen gaan de spectrale profielen van de centra van de valplekken lijken op de spectrale profielen van braak liggende grond.

In Figuren 2a en 2b zijn biplots weergegeven van principale componenten analyse (PCA) op respectievelijk de centra en de randen van de valplekken (gemiddelde spectra, gecorrigeerd voor gezond gewas). Voor beide principale componenten analyses geldt dat met twee principale componenten (PC) een zeer groot gedeelte van de variatie in de spectrale gegevens verklaard wordt. Voor de centra van de valplekken is dat 93% (85.44 + 7.61), voor de randen van de valplekken is dat bijna 97% (92.82 + 4.03). In beide biplots is te zien dat de variatie in de spectrale kanalen tussen 760 en 870 nm ongeveer hetzelfde is (ze vormen een cluster in de biplot). Hetzelfde geldt in iets mindere mate ook voor de kanalen met golflengte kleiner dan 710 nm. De hogere golflengte gebieden liggen wat meer verspreid in de tweedimensionale PC-ruimte. Wat in beide biplots erg opvalt is dat de verschillende categorieën valplekken niet in verschillende delen van de PC-ruimte liggen, en dat met name de valplekken veroorzaakt door structuurbederf (codes 3-9) erg verspreid over de ruimte liggen. Aangezien met 2 PC's zo'n 95% van de variatie in de spectrale gegevens verklaard wordt, geven de principale componenten analyses al een indicatie van de moeilijkheden die verwacht mogen worden bij de classificatie van valplekken op basis van cropscan metingen. Mogelijk is wel dat deze moeilijkheden van minder belang zijn indien niet alle drie de klassen in één model onderscheiden hoeven te worden.

PLS-DA

De resultaten van PLS-DA zijn weergegeven in de Tabellen 1a t/m 1e, waarbij in elk van de tabellen modellen voor het onderscheiden van verschillende klassen zijn gemaakt. Voor de respectievelijke tabellen is dat het onderscheiden van:

a) de drie verschillende categorieën onderling,
b) categorie 1+2 gezamenlijk versus de derde categorie (structuurbederf),
c) geeft de resultaten van onderscheiden van categorie 1 (aardappelmoeheid) ten opzicht van 2+3 gezamenlijk,
d) categorie 1 (aardappelmoeheid) vs. categorie 3 (structuurbederf),
e) categorie 1 vs. categorie 2 (levende nematoden).

Tijdens de selectie van het aantal PLS componenten bleek er vaak niet een echt minimum te onderscheiden te zijn, of zich meerdere locale minima voor te doen. Bovendien bleken de twee criteria (LOO-SEP en percentage correct geclassificeerde objecten op basis van de LOO berekeningen), niet altijd consistent. In de meeste gevallen is er gekozen voor een zo spaarzaam mogelijk model, dat wil zeggen met het kleinst aantal PLS componenten, waarbij het percentage correct criterium zwaarder is gewogen in de selectie.

De leave-one-out resultaten zijn ook weergegeven, aangezien voor de SEP en het percentage correct geclassificeerd geldt dat naarmate het aantal PLS componenten toeneemt op deze criteria beter gescoord zal worden. Voor de leave-one-out criteria geldt dit niet (opmerking: de LOO SEP kan groter worden dan SST).

Uit de Tabellen 1a t/m 1e blijkt dat met name de classificatie van categorie 1 ten opzichte van de beiden andere categorieën, of één van de andere categorie redelijk goed gaat. De percentages goed geclassificeerde valplekken op basis van kruisvalidatie bedraagt ongeveer 75% of hoger. Gegeven het feit dat met name de valplekken in categorie 3 nogal divers van aard zijn, en het aantal bemonsterde valplekken in elke categorie ook niet groot is kan dit gezien worden als een positief resultaat.

Tabel 1a. Resultaten PLS-DA, categorie 1 vs. 2 vs. 3 (totale kwadratensom SST = 192).

	Aantal PLS componenten	LOO SEP	SEP	LOO perc. correct	perc. correct
Centrum	7	185	105	66	75
Centrum (gecorr. gezond)	2	151	135	59	62
Rand	6	141	100	59	69
Rand (gecorr. gezond)	4	152	119	67	75

Tabel 1b. Resultaten PLS-DA, categorie 1+ 2 vs. 3 (totale kwadratensom SST = 128).

	Aantal PLS componenten	LOO SEP	SEP	LOO perc. correct	perc. correct
Centrum	2	132	104	58	66
Centrum (gecorr. gezond)	2	116	99	69	76
Rand	4	124	93	70	72
Rand (gecorr. gezond)	7	114	71	75	83

Tabel 1c. Resultaten PLS-DA, categorie 1 vs. 2 + 3 (totale kwadratensom SST = 128).

	Aantal PLS componenten	LOO SEP	SEP	LOO perc. correct	perc. correct
Centrum	7	96	61	83	86
Centrum (gecorr. gezond)	3	101	76	73	85
Rand	11	69	39	86	95
Rand (gecorr. gezond)	3	104	80	80	84

Tabel 1d. Resultaten PLS-DA, categorie 1 vs. 3 (totale kwadratensom SST = 102).

	Aantal PLS componenten	LOO SEP	SEP	LOO perc. correct	perc. correct
Centrum	7	92	50	80	84
Centrum (gecorr. gezond)	2	86	73	74	80
Rand	3	73	55	74	76
Rand (gecorr. gezond)	3	85	60	78	82

Tabel 1e. Resultaten PLS-DA, categorie 1 vs. 2 (totale kwadratensom SST = 80).

	Aantal PLS componenten	LOO SEP	SEP	LOO perc. correct	perc. correct
Centrum	11	22	10	95	100
Centrum (gecorr. gezond)	3	44	33	80	88
Rand	5	30	14	98	100
Rand (gecorr. gezond)	3	38	29	82	90
Centrum - rand	4	66	44	75	80

SIMCA

De resultaten van de SIMCA analyses zijn weergegeven in de Tabellen 2a en 2b. Alleen de resultaten voor het onderscheiden van de drie categorieën onderling en de resultaten voor het onderscheiden van categorie 1 en 2 zijn opgenomen in deze nota. Het toekennen van lidmaatschap van een valplek aan een bepaalde klasse wordt mede bepaald door de spreiding binnen een klasse. Er worden grenzen berekend van de afzonderlijke klassen (in de vorm van 95% limieten van T^2 en Q-waarden, zie genoemde referenties voor een uitgebreide beschrijving hiervan). Aangezien de valplekken in categorie 3 zeer divers zijn, levert dit een grote binnenklasse spreiding op. Als gevolg hiervan worden bij voorspellingen veel van de valplekken uit andere categorieën in categorie 3 geclassificeerd (zie als voorbeeld de resultaten in Tabel 2a). Over het algemeen worden alle objecten uit categorie 3 goed geclassificeerd, echter ook voor categorieën 1 en 2 worden de meeste objecten in categorie 3 geclassificeerd. Bijvoorbeeld, voor centrum is dat 13 van de 27 categorie 1 objecten (de andere 14 worden goed geclassificeerd), en 5 van de 13 categorie 2 objecten (8 goed geclassificeerd). De resultaten van de SIMCA analyses zijn alleen goed indien deze categorie 3 valplekken buiten de analyses wordt gehouden (zie Tabel 2b).

Tabel 2a. Resultaten SIMCA, categorie 1 vs. 2 vs. 3.

	PCA comp. per categorie	perc. correct
Centrum	5,5,3	70
Centrum (gecorr. gezond)	7,4,3	48
Rand	5,3,2	62
Rand (gecorr. gezond)	5,3,2	56

Tabel 2b. Resultaten SIMCA, categorie 1 vs. 2.

	PCA comp. per categorie	perc. correct
Centrum	5,5	98
Centrum (gecorr. gezond)	7,4	88
Rand	5,3	92
Rand (gecorr. gezond)	5,3	88
Centrum - rand	7,4	88

Aanvullende analyses

Naar aanleiding van de resultaten beschreven in voorgaande paragrafen, zijn enkele aanvullende analyses gedaan met behulp van niet-lineaire methoden. Kwadratische PLS-DA, waarbij de spectrale dataset werd uitgebreid door kwadratische termen op te nemen (van de spectrale variabelen), leverde geen verbetering van de resultaten op. Een kleine verkennende studie naar de mogelijkheden van toepassing van een feedforward neuraal netwerk levert voorlopig nog geen verbetering op (over het algemeen zelfs slechtere resultaten). Voorlopig is er daarom voor gekozen om dit niet verder uit te diepen.

Conclusie

De spectrale gegevens zoals ze nu beschikbaar zijn, blijken niet discriminerend genoeg te zijn om een onderscheid te kunnen maken tussen de drie gedefinieerde valplek categorieën. Dit wordt voor het overgrote deel veroorzaakt door de grote spreiding die er is binnen de categorie valplekken veroorzaakt door structuurbederf. De resultaten voor het onderscheid van categorie 1 valplekken ten opzichte van de andere categorieën zijn wel goed.
De belangrijkste aanbevelingen voor verder onderzoek zijn dan ook:

1. De dataset uitbreiden met meer valplekken uit categorie 3, en dan het aantal categorieën uitbreiden zodat de variatie binnen categorieën kleiner wordt, door onderscheid te maken tussen de verschillende subklassen,
2. Extra variabelen meten (welke?) waarin het onderscheid tussen de verschillende klassen beter tot uitdrukking komt.

Indien aan punten 1. en/of 2. tegemoet gekomen kan worden lijkt het zinvol om de analyses te herhalen, en daarbij te concentreren op PLS-DA en neurale netwerken.

Referenties

Sjöström, M., S. Wold & B. Söderström.
PLS discriminant plots. In: Pattern recognition in practice II. E.S. Gelsema and L.N. Kanal, Eds., Elsevier, Amsterdam, 1986.

Wold, S.
Pattern recognition by means of disjoint principal component models. *Patt. Recog.* 8, 127-139 (1976).

Lavine, B.K.
Chemometrics: Fundamental review. *Anal. Chem.* 72, 91R-98R (2000a).

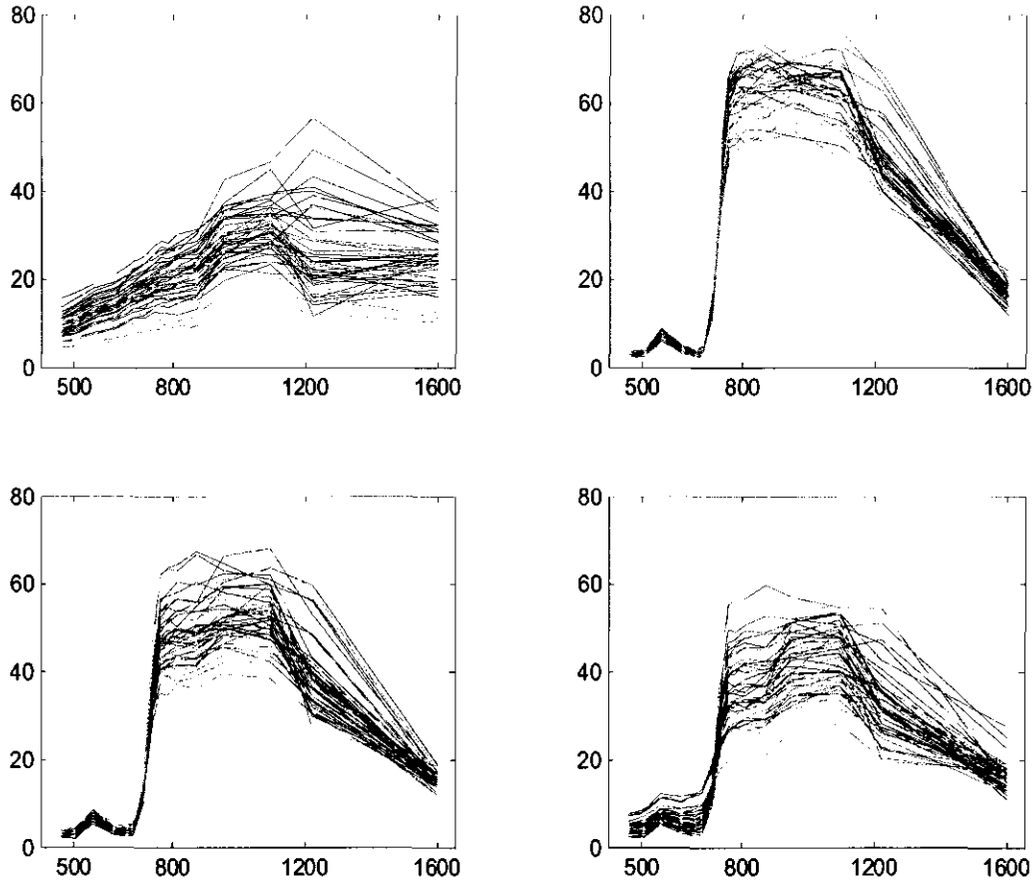
Matlab, versie 6.1, release 12, The Mathworks, Inc., 2000

PLS_Toolbox, versie 2.1, Eigenvector Research, Inc. Manson, WA, 2000.

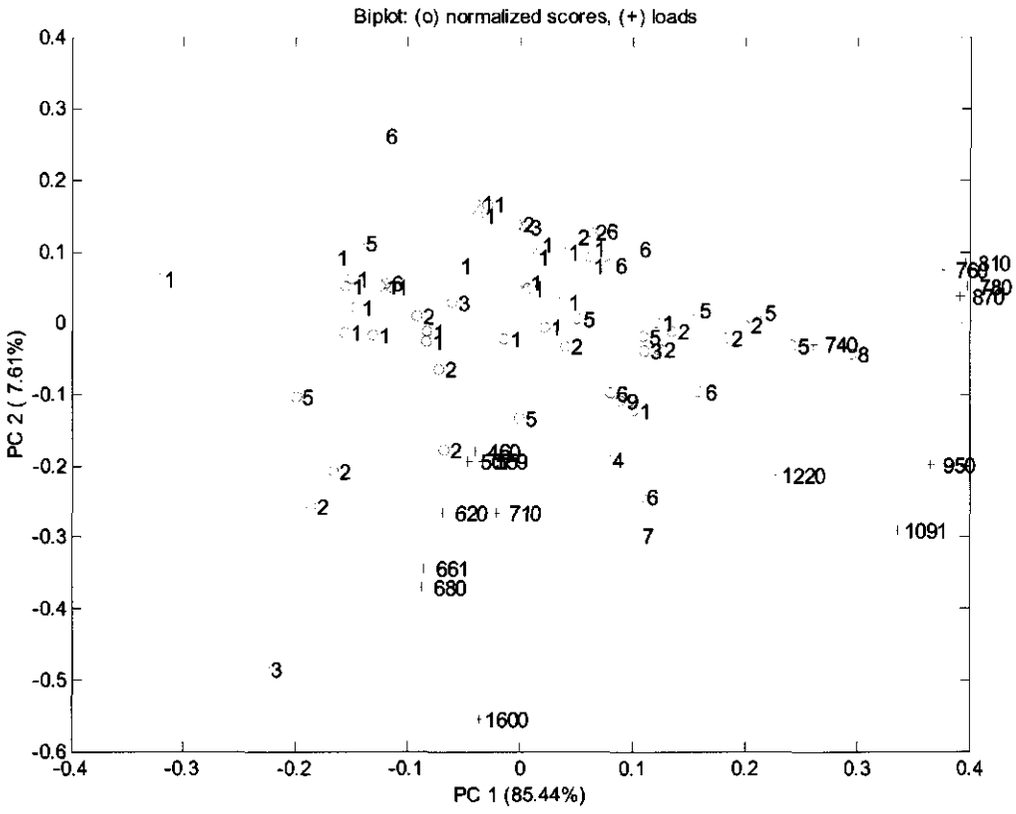
Bijlage I.
Figuren

Figuur 1a. *Afzonderlijke spectra per veld 2000.*
rood = gezond , geel = rand valplek, groen = centrum valplek, blauw = braak liggende grond.

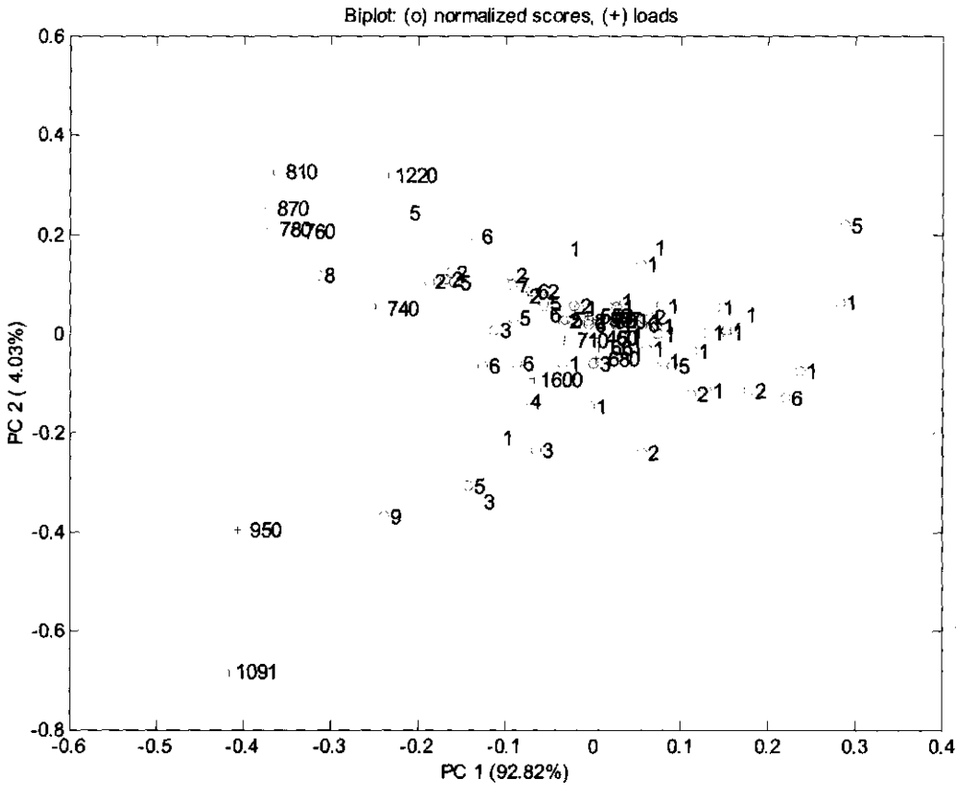
Figuur 1b. *Afzonderlijke spectra per veld 2001.*
rood = gezond , geel = rand valplek, groen = centrum valplek, blauw = braak liggende grond



Figuur 1c. *Gemiddelde spectra alle valplekken data. lb: braak, rb: gezond, lo: rand, ro: centrum valplek. rood = aardappelmoeheid (categorie 1), blauw = nematoden (categorie 2), groen = structuurbederf (categorie 3)*



Figuur 2a. *Biplot Principale componenten analyse centrum valplekken (gecorrigeerd voor gezond gewas) , codering van plekken (rode rondjes) komt overeen met beschrijving op pagina 1 van deze notitie, de blauwe kruisjes geven de spectrale variabelen weer.*



Figuur 2b. *Biplot Principale componenten analyse randen valplekken (gecorrigeerd voor gezond gewas), codering van plekken (rode rondjes) komt overeen met beschrijving op pagina 1 van deze notitie, de blauwe kruisjes geven de spectrale variabelen weer.*