Opbouw Methodiek Prijsbepaling Hout

Andreas Demey, Lander Baeten en Kris Verheyen
(ForNaLab, Universiteit Gent)

Agentschap voor Natuur en Bos

inverde
forum voor groenexpertise

Colofon

Deze studie werd uitgevoerd door ForNaLab (Universiteit Gent) in opdracht van het Agentschap voor Natuur en Bos en van Inverde.
Dit rapport is een gezamenlijke uitgave van het Agentschap voor Natuur en Bos en van Inverde
Koning Albert II-laan 20 bus 8, 1000 Brussel
www.natuurenbos.be – www.inverde.be
Contact: info@inverde.be

Dit rapport is opgemaakt in het kader van het KOBE-project. KOBE staat voor KennisOndersteuning bij Beheer en Economie van natuur-, groen- en bosdomeinen. KOBE is een samenwerkingsproject tussen het Agentschap voor Natuur en Bos en Inverde. Dit rapport is een werkdocument, en weerspiegelt niet noodzakelijk de standpunten of de werking van het Agentschap voor Natuur en Bos en Inverde.

Auteurs: Andreas Demey, Lander Baeten en Kris Verheyen (ForNaLab, Universiteit Gent)

Uitgave: augustus / 2013

Dit rapport is ook beschikbaar op het ANB-intranet (http://teamlne.vlaanderen.be/anb/intranet/Paginas/default.aspx)

Inhoudsopgave

1 Probleem- en doelstelling

Het Agentschap voor Natuur en Bos (ANB) verkoopt jaarlijks een 200.000 m³ hout. Het wordt verkocht via een openbare houtverkoop, waarbij loten op stam worden aangeboden en verkocht via de methode van opbod, afbod of inschrijving. De hoogste bieder krijgt het lot toegewezen, op voorwaarde dat de vooropgestelde schattingsprijs (mits een mogelijke afwijking) wordt gehaald. Aangezien de loten heel gevarieerd zijn qua samenstelling, zowel naar soorten, omtreksklassen als kwaliteiten en aangezien de prijszetting gebeurt per lot is het *heel moeilijk om voeling te krijgen met de uiteindelijk geboden prijzen per boomsoort, sortiment of kwaliteitsklasse*. Momenteel bestaat er nog geen goede methodiek met betrekking tot het inschatten van de houtprijzen.
Houtprijzen worden beïnvloed door kenmerken die direct gekoppeld zijn aan het lot (volume, boomsoort(en), kwaliteit hout, dimensies, soort kap, moeilijkheid exploitatie) en door externe factoren (situatie op de houtmarkt, nabijheid van een verwerkend bedrijf, …). Het kwantificeren van de invloed van deze externe factoren is erg complex en vereist de verzameling van veel bijkomende gegevens. De gegevens van houtverkopen die ANB de afgelopen jaren verzameld heeft, laten echter wel toe om meer inzicht te krijgen in het belang van kenmerken die direct gekoppeld zijn aan het lot (de lotkenmerken).

De *algemene doelstelling* van de opdracht is ANB meer inzicht te verschaffen in de waarde-inschatting van het hout dat zij aanbiedt. De *specifieke doelstellingen* zijn:

(1) de impact van verschillende lotkenmerken op de verkoopswaarde kwantificeren door statistische analyse van de verkoopgegevens van de voorbije 4-5 jaar, ter beschikking gesteld door ANB;

(2) de variabiliteit van prijzen die niet door lotkenmerken verklaard worden inschatten (bv. tussen jaren, regio's);

(3) een voorstel aanleveren voor een eerste eenvoudige methodiek ter bepaling van de schattingsprijs van hout dat door ANB wordt verkocht (predictie).

2 Materiaal en methode

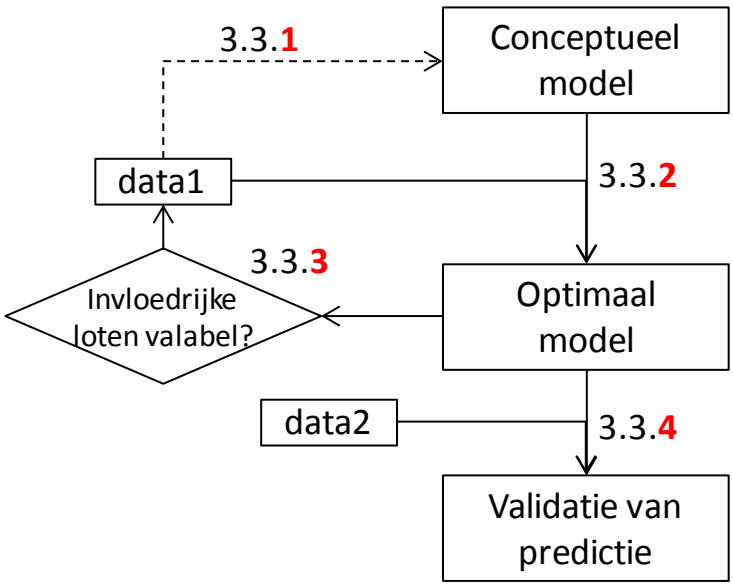
2.1 Beschikbare gegevens

ANB leverde een datafile (data1, Excel) aan met de gegevens van de houtverkoop van 2008 tot medio 2012, bestaande uit 2450 loten. Deze datafile bevat volgende gegevens: dienstjaar, provincie, regio, type bos (domeinbos of ander openbaar bos), naam bos, lotnummer, boomsoort, omtrekklasse, hoogte, volume *per boomsoort en omtrekklasse*, prijs *per lot*, FSC gecertificeerd of niet. Halverwege de loop van het project werd een tweede datafile (data2) aangeleverd met de meest recente verkoopsgegevens. Deze datafile bevat dezelfde gegevens als data1, voor de loten verkocht in 2013 nog aangevuld met de variabelen 'Kwaliteit Type' (score voor houtkwaliteit) en 'Exploitatiegraad' (score voor exploitatie omstandigheden).

Voor de verdere analyse is het zeer belangrijk zich te realiseren dat de gegevens in de ANB-datafile niet onmiddellijk toelaten de prijs per soort en/of omtrekklasse af te leiden omdat enkel de prijs per lot gekend is en een lot bijna steeds uit meerdere boomsoorten en/of omtrekklassen bestaat.

2.2 Statistische analyse van de verkoopgegevens

De statistische analyse begint met de controle en vereenvoudiging van de dataset data1 (3.1). Daarna wordt, ter voorbereiding van de analyse, meer inzicht verkregen in de datastructuur (3.2). In de eigenlijke analyse (3.3) wordt dan, op basis van de verworven inzichten, een conceptueel model opgebouwd (3.3.1). Dit model bevat het maximaal aantal parameters dat overwogen wordt. Via een optimalisatieprocedure (3.3.2), wordt het model terug vereenvoudigd tot het optimale aantal parameters voor data1. Vervolgens worden invloedrijke loten in kaart gebracht (3.3.3). Wanneer deze niet valabel zijn, wordt het model opnieuw gefit zonder de desbetreffende loten. In een laatste stap (3.3.4) worden de predicties van het model gevalideerd op basis van een nieuwe set data, data2. Het verschil tussen predicties en de werkelijke verkoopsprijzen wordt tenslotte in verband gebracht met de nieuwe variabelen 'Exploitatiegraad' en 'Kwaliteit type' voor de loten uit het voorjaar 2013.



Figuur 1: Flowchart die gevolgd werd bij het modelleren

3 Analyse

3.1 Controle en vereenvoudiging van de dataset

3.1.1 Controle op fouten

De dataset (data1) omvat 2450 loten verkocht tussen 2008 en 2012. Deze werden gecontroleerd op typefouten of andere mogelijke foute invoer en aangepast of verwijderd waar nodig (Tabel 1). Loten waarvan de som van de deelvolumes voor omtrek-boomsoort combinaties niet gelijk is aan totaal volume werden niet meegenomen in de analyse. Het gaat over de loten 3-2010/vag-012, 6-2009/dom-002, 11-2012/win-014, 3-2012/mor-003 en 8-2008/hev-121.

Tabel 1: Logboek van de initiële correcties van de dataset

datum	gewijzigd
6/05/2013	spaties verwijderd in Omtrekklasse om ze in zelfde format te krijgen
6/05/2013	Omtrekklasse='0' (fout?). Geen fout, maar omtrekklasse is onbekend.
6/05/2013	FSC levels altijd 'FALSE' or 'TRUE' (waren soms 'WAAR' of 'ONWAAR')
6/05/2013	lotvolumes berekend voor nieuwe data (dienstjaar 2013)
6/05/2013	kolommen 'Boomsrt_recl', 'Omtrek_ondergrens', 'Omtr_recl', 'VolumeFractie' en 'PrijsPerM3' toegevoegd
6/05/2013	lot zonder volume verwijderd
6/05/2013	13-2010/hal-021 --> 13-2010/hal-021a (2010) en 13-2010/hal-021b (2011)
6/05/2013	15-2009/bki-001b --> 15-2009/bki-001a (2009) en 15-2009/bki-001b (2010)
8/05/2013	14-2012/slo-001 --> volledig lot aangeduid als FSC (in opmerking staat nog originele waarden)
8/05/2013	'Openbare verkoop' en 'Openbare Verkoop' --> 'Openbare Verkoop'
14/05/2013	3-2011/mil-061 en 3-2011/ede-062: 'Ander openbaar bos' --> 'Ander Openbaar Bos'
14/05/2013	Naamgeving provincies en Regio's consitent gemaakt
14/05/2013	Namen van kopers gecodeerd (anonieme verwerking)
14/05/2013	Controle of lotvolume gelijk is aan som van deelvolumes van omtrek-boomsoort combinaties

3.1.2 Controle op outliers

Via boxplots en scatterplots werd gecontroleerd voor sterk afwijkende observaties ('outliers'). Deze kunnen immers te wijten zijn aan foutieve datainvoer (bv. een nul teveel of fout geplaatste komma) en dienen op validiteit gecontroleerd te worden. Twee loten, 12-2009/llp-001 en 3-2010/vag-127, hadden een prijs per m³ hoger dan 90, in combinatie met een uitzonderlijk klein volume voor dit prijsniveau (17m³ en 0.74 m³). Een ander lot, 13-2012/hal-022, had een zeer kleine prijs per m³ (0.007). Deze loten werden uitgesloten voor analyse. Ook invloedrijke observaties – loten die de prijsschattingen sterk beïnvloeden – worden best nagekeken. Het kan bijvoorbeeld zijn dat een lot verkocht werd aan een afwijkende prijs omwille van exploitatievoorwaarden of kwaliteit van de stammen. Na het fitten van het model werden deze invloedrijke loten in kaart gebracht (zie deel 3.3.3). Na het eventueel verwijderen van invloedrijke loten (op basis van extra gegevens over het betreffende lot), werd het model opnieuw gefit.

3.1.3 Vereenvoudiging van de dataset

De beschikbare data omvat een vijftigtal houtsoorten onderverdeeld in omtrekklassen per 10 cm. Dit is te fijn om betrouwbare prijsschattingen te kunnen maken per categorie, omdat in dergelijk model te veel parameters geschat moeten worden wat een grote onzekerheid in de prijsschatting oplevert. In samenspraak met de stuurgroep werden minder courante soorten ondergebracht in de restcategoriën 'ander loof' (Aloof) en 'ander naald' (Anaald) (Tabel 2). De 14 overgehouden houtsoorten zijn Amerikaanse eik (Aeik), berk, beuk, boskers, Corsicaanse den (CorsDen), douglas, es, esdoorn, fijnspar, grove den (Gden), inlandse eik (InlEik), larix, populier en tamme kastanje (Tkastanje).

Tabel 2: Restcategorieën voor 'ander loof' (Aloof) en 'ander naald' (Anaald). De namen van de houtsoorten zijn overgenomen uit de oorspronkelijke dataset.

Aloof		Anaald	
Abeel	Robinia	ander naaldhout	xxxn
Am. Vogelkers	Veldesdoorn	Californische cipres	Cipres sp.
ander loofhout	Vlier	Ceder	Den sp.
Gelderse roos	Vogelkers (Eur.)	Cypres	Sequoia
Haagbeuk	Vuilboom	Hemlockspar	Spar sp.
Hazelaar	Walnoot	Levensboom	
Lijsterbes	Wilde appel	Oostenrijkse den	
Linde	Wilg	Reuzenzilverspar	
Meidoorn	Witte els	Sitkaspar	
Moeraseik	Zure kers	Spar	
Olm	Zwarte els	Taxus	
Paardekastanje	Zwarte populier	Weymouthden	
Plataan	xxxl	Zeeden	
Ratelpopulier	Wilg sp.	Zilverspar	

In samenspraak met de stuurgroep werden de omtrekklassen vereenvoudigd naar zeven klassen per 50 cm voor loofhout en zes klassen per 40 cm voor naaldhout (Tabel 3). Afhankelijk van de volumeverdeling van omtrekklassen per boomsoort kwamen echter niet alle zes of zeven klassen voor bij elke boomsoort (Figuur 2).

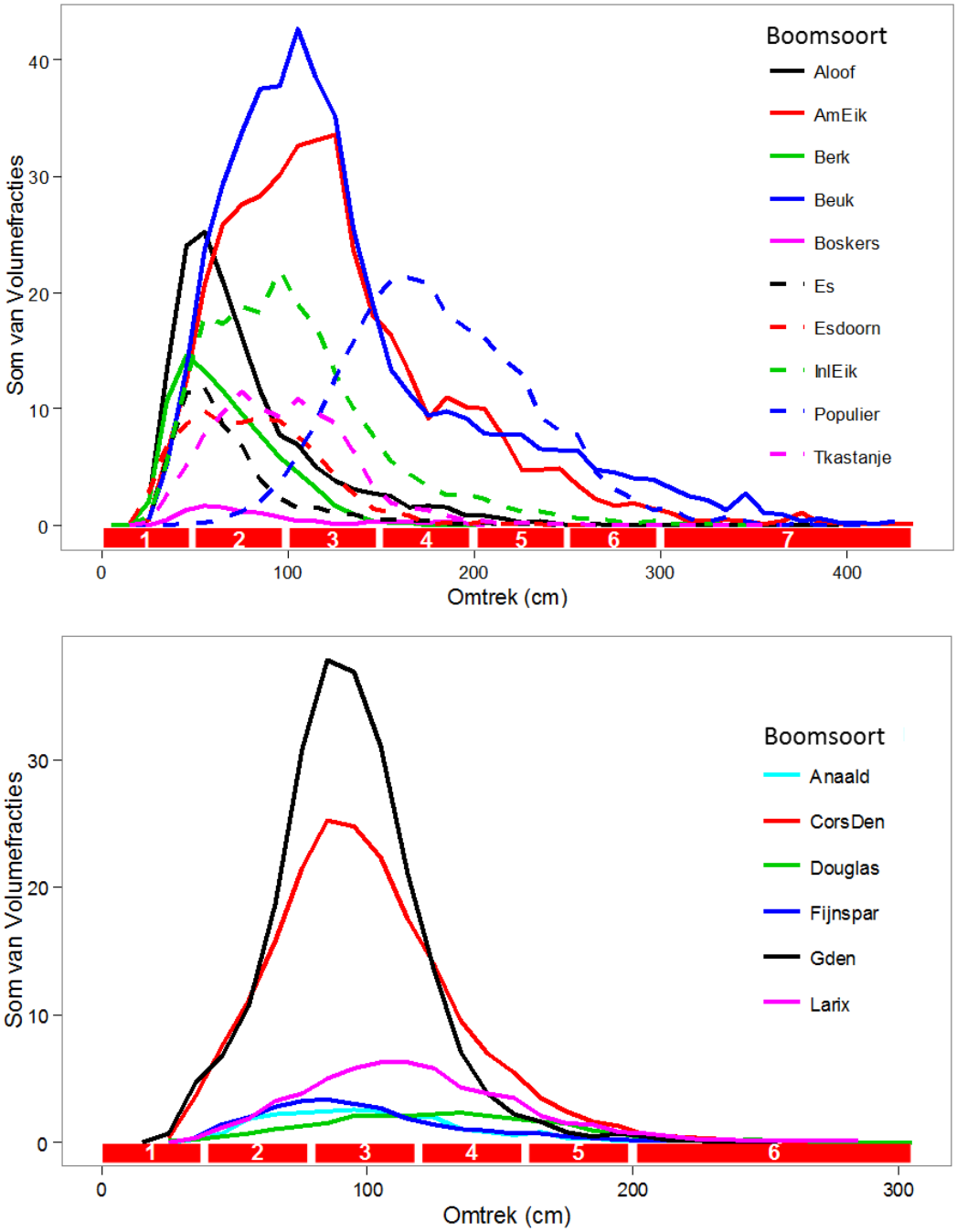
Tabel 3: Herklassering van omtrekklassen (cm) voor loof- en naaldhout in 7 categorieën.

Klasse	Loof	Naald
1	0 - 49	0 - 39
2	50 - 99	40 - 79
3	100 - 149	80 - 119
4	150 - 199	120 - 159
5	200 - 249	160 - 199
6	250 - 299	200 - 239
7	> 300	> 240

De variabele op lotniveau 'Soort Verkoop' heeft als categorieën 'Eigen Verkoop', 'Openbare verkoop' en 'Openbare Herverkoop'. 'Eigen verkoop' (26 loten) omvatten loten die door openbare besturen aan eigen werknemers werden verkocht dus niet via een openbare verkoop. Deze loten werden niet weerhouden voor analyse. Tussen de categorieën 'Openbare Verkoop' (2405 loten) en 'Openbare Herverkoop' (6 loten) werd verder geen onderscheid gemaakt daar het telkens om een openbare verkoop gaat.
De variabele 'Dienstjaar' geeft het jaar van velling weer. Het eigenlijke moment van verkoop is in de meeste gevallen het najaar van het jaar voordien. Daarom werd 'Dienstjaar' vervangen door de variabele 'Jaar' (het jaar van verkoop), gelijk aan dienstjaar min één. De prijzen werden omgerekend naar het prijsniveau in 2008 gebruik makend van de consumptieprijsindex (Tabel 4).

Tabel 4: Herrekening van de prijs naar het prijsniveau van 2008 ($Prijs_{2008}$) op basis van de Consumtieprijsindex.

Jaar	$Index_{2004}$	$Prijs_{2008}$
2008	111.32	Prijs
2009	111.26	Prijs/111.26*111.32
2010	113.69	Prijs/113.69*111.32
2011	117.71	Prijs/117.71*111.32
2012	121.05	Prijs/121.05*111.32



Figuur 2: Herklassificatie van de omtrekklassen (cm) per boomsoort. De curves geven de volumeverdeling – de som van de volumefracties per boomsoort-omtrekklasse combinatie over alle loten heen – in de originele data weer. De zeven (loof) of zes (naald) nieuwe categorieën zijn aangeduid met genummerde rode blokjes.

3.1.4 Herstructureren van de dataset

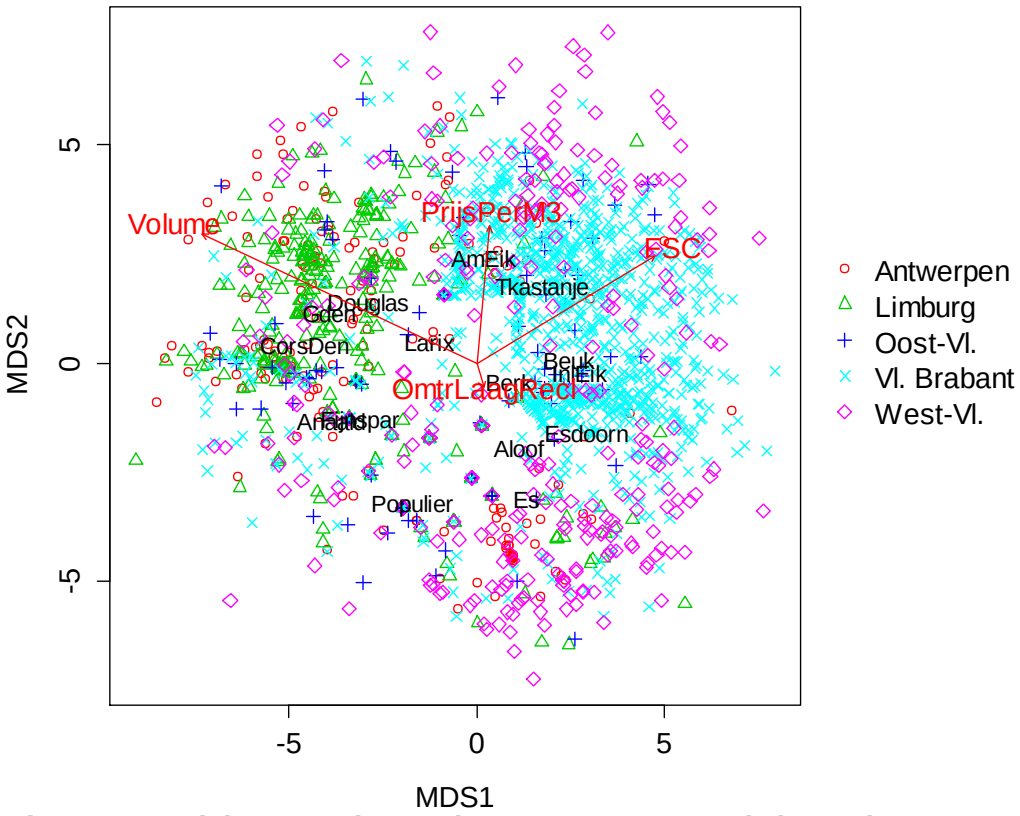
De oorspronkelijke dataset is opgesteld in 'lang' formaat. Elke record (lijn) geeft data voor een bepaalde boomsoort-diameterklasse combinatie binnen een lot weer. Omdat we slechts een prijs op lotniveau hebben, dient de structuur van de data zo te worden aangepast dat elke lijn een lot vertegenwoordigd ('breed' formaat). Dit kan via een kruistabel. Elke boomsoort-omtrekklasse combinatie vormt een nieuwe variabele (kolommen in de data) met waarden die aangeven wat de volumefractie binnen het lot is. Ook de andere weerhouden variabelen dienen op lotniveau te zijn. Terwijl dat voor de meeste variabelen het geval is (Totaal Volume, Prijs, Provincie, Regio, FSC, Koper, …), zijn er ook die variëren binnen een lot (Type Kap, LoofNaald). Daarom moeten deze worden herrekend (Tabel 5).

Tabel 5: Een lot wordt in twee categorieën van de variabelen 'LoofNaald' en 'Type Kap' ondergebracht wanneer minstens 60% van het volume in het lot tot die categorie behoort. De andere loten worden toegewezen aan een restcategorie

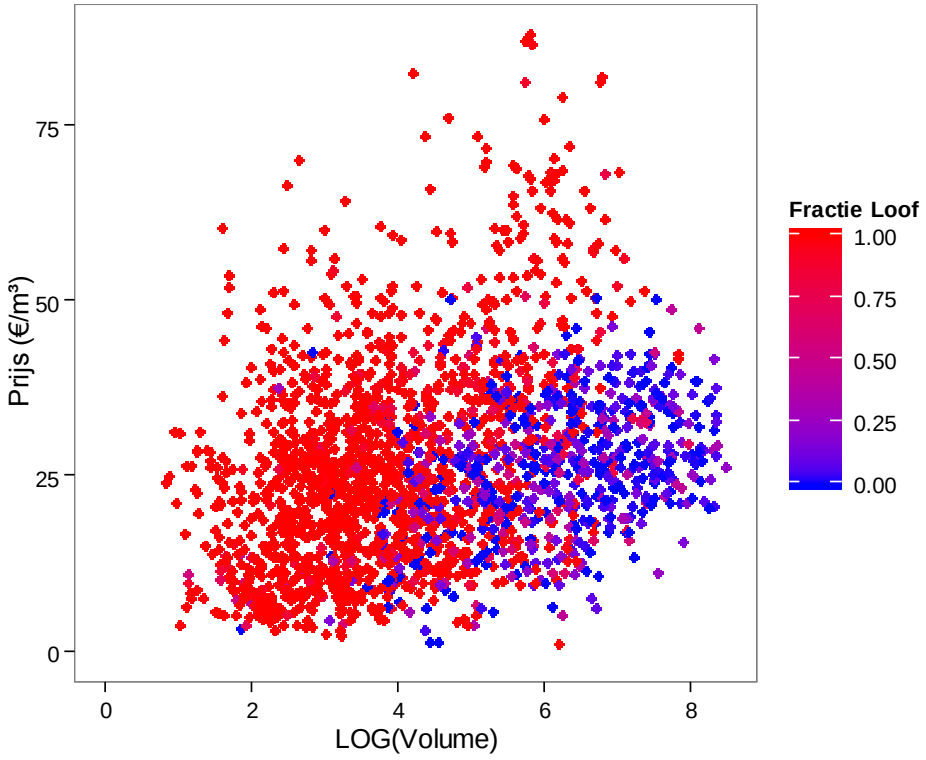
LoofNaald		Type Kap	
oud	nieuw	oud	nieuw
>60% Loof	Loof	>60% Dunning	Dunning
>60% Naald	Naald	>60% Kaalkap	Kaalkap
<60% Loof en <60% Naald	Gemengd	<60% Dunning en <60% Kaalkap	Andere

3.2 Verkenning van de data

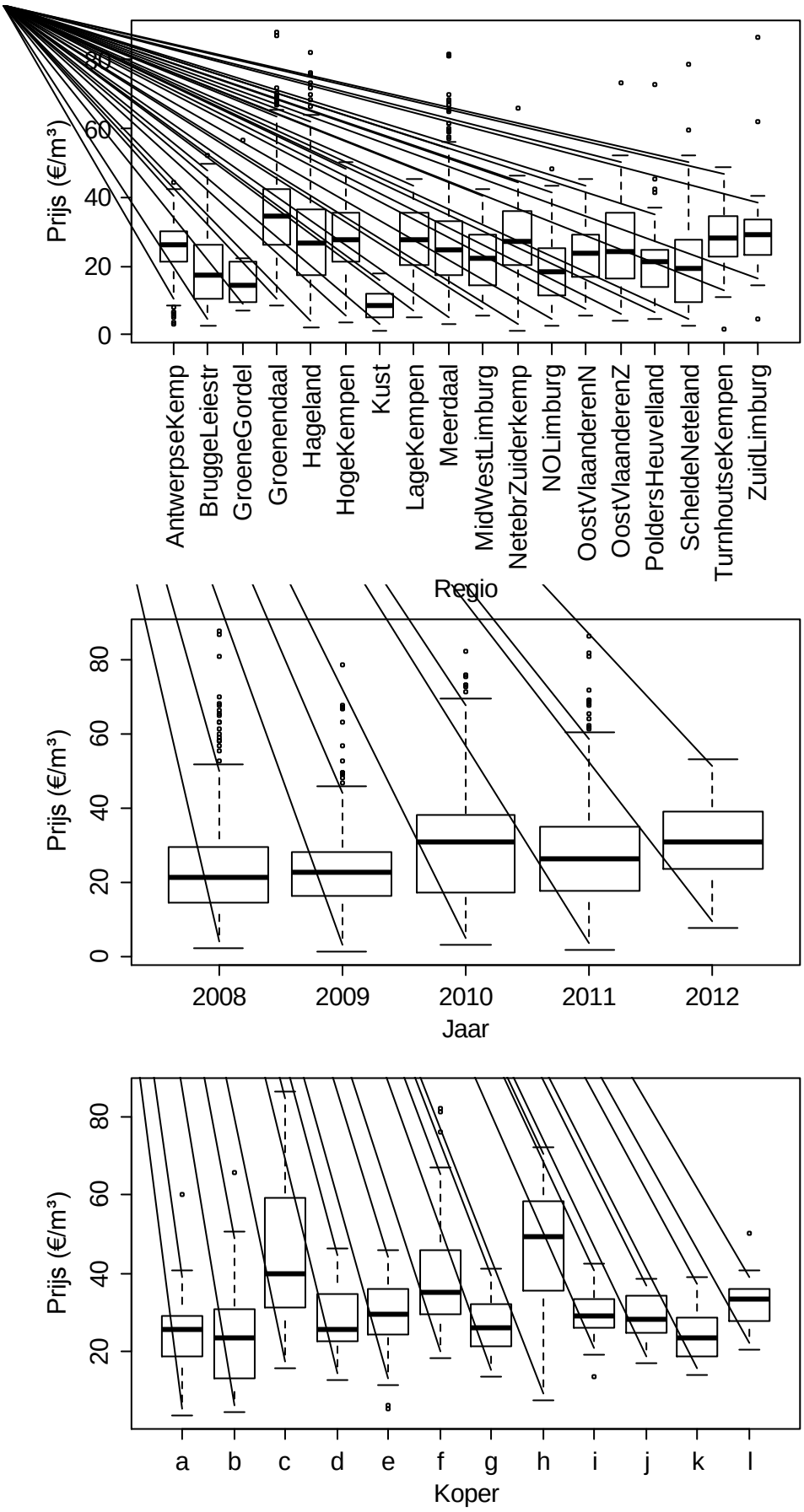
Om inzicht te verwerven in de data werden eerst enkele exploratieve figuren gemaakt. In Figuur 3 zien we een NMDS ordinatie op basis van een euclidische afstand van de boomsoortsamenstelling van de loten. Deze analyse plaatst loten (punten in de figuur) die sterk op elkaar lijken zo dicht mogelijk bij elkaar, terwijl sterk verschillende loten ver van elkaar staan. De locatie van de boomsoortnamen op de figuur geeft een indicatie in welke cluster van loten deze soort sterk vertegenwoordigd is. Ten slotte werden enkele andere variabelen *a posteriori* gerelateerd met de variatie in boomsoortsamenstelling van de loten. Vooral naaldhoutloten halen hoge volumes, het loofhout uit Vlaams Brabant is vaker FSC gelabeld en de hoogste prijzen per m³ vind je zowel voor naaldhoutloten, meestal uit Limburg, en voor loofhoutloten, meestal uit Vlaams Brabant. De gemiddelde omtrek van bomen in een lot is niet of nauwelijks gerelateerd aan de boomsoortsamenstelling.
Een algemeen stijgende trend in prijs per m³ is zichtbaar met stijgend lotvolume (Figuur 4). Deze trend is duidelijk verschillend voor loten met overwegend loof- of naaldhout. Daarom werd, in samenspraak met de stuurgroep, beslist om de interactie tussen volume en de factor 'LoofNaald' in het model op te nemen. Zo wordt voor Loof, Naald en Gemengde loten een ander volume-effect op de kubieke meterprijs geschat. Dit wil zeggen dat men een meerprijs wil betalen voor grotere naaldhoutloten, maar niet voor grotere loofhoutloten (zie deel 3.4)
De gemiddelde prijs per m³, is vrij variabel tussen verschillende regio's en jaren (Figuur 5, boven en midden). In regio Kust, bijvoorbeeld, liggen de prijzen erg laag, terwijl we de hoogste prijzen terugvinden in regio Groenendaal. We zien de gemiddelde prijs ook wat stijgen over de jaren heen (in 2011 was er een kleine terugval), maar deze stijging is relatief beperkt gezien de grote variatie binnen jaren. Ook wat kopers betreft (Figuur 5, onder) zien we dat er heel wat variatie is. Koper c en h, bijvoorbeeld, betalen een prijs per m³ die een stuk boven het gemiddelde ligt. De gemiddelde prijs voor FSC loten is iets hoger dan deze voor niet-FSC (Figuur 6, links), maar dit verschil is niet significant. Loten waarvan meer dan 60% van het volume via kaalkappen geëxploiteerd is, zijn gemiddeld iets goedkoper dan deze hoofdzakelijk via dunningen geëxploiteerd (Figuur 6, rechts). De prijsvariatie tussen regio's, jaren, type kap, al dan niet FSC loten en kopers is telkens veel minder uitgesproken dan de variatie binnen deze categoriën. Toch kan het in rekening brengen van deze variabelen als 'random effecten' (zie deel 3.3) een deel residuele variatie opvangen. Deze random effecten zijn wellicht deels gerelateerd aan variabelen die nu nog niet in data1 zitten (bv. 'Kwaliteit Type' en 'Exploitatiegraad'), maar kan evengoed onverklaard blijven.
Omdat loten met populier meestal geen of nauwelijks andere soorten omvatten, werd besloten twee aparte analyses uit te voeren: één voor alle loten met maximum 5% populier, en één voor alle loten met minimum 95% populier. In de eerste analyse werden alleen boomsoorten anders dan populier in het model opgenomen, in de tweede analyse enkel populier.



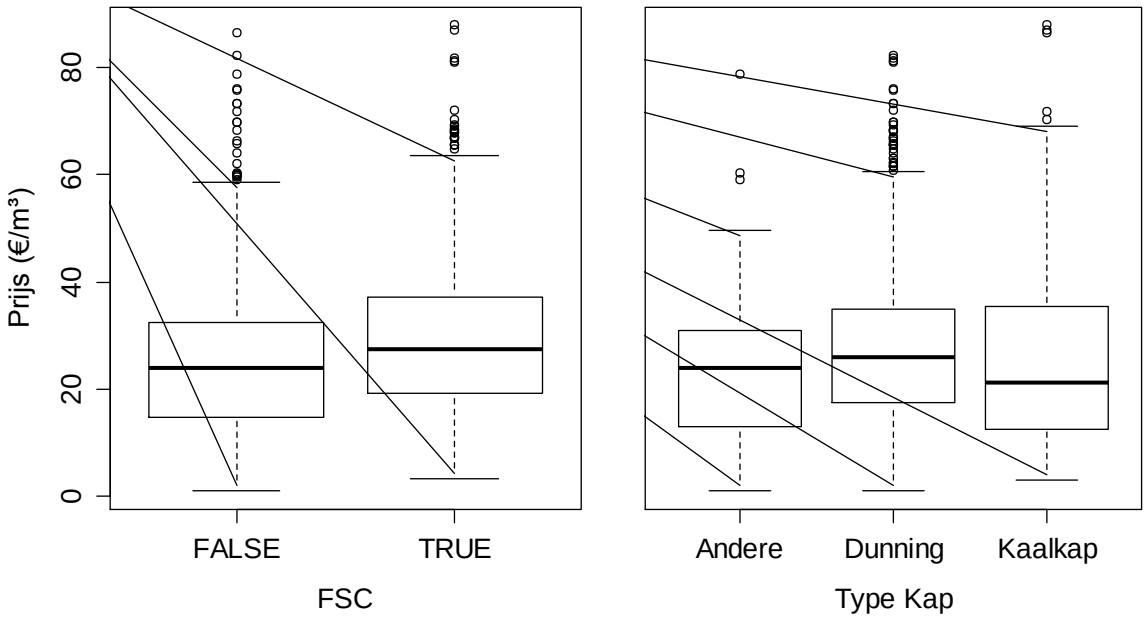
Figuur 3: Triplot van loten, boomsoorten en enkele andere lotkenmerken (volume, prijs per m³, FSC of niet en gemiddelde omtrek) verkregen door een NMDS ordinatie. De afstand tussen loten in de figuur geeft weer hoe verschillend ze zijn in samenstelling van boomsoorten. Verschillende kleuren geven de provincies weer.



Figuur 4: Prijs (prijsniveau 2008) in funtie van volume (log-getransformeerd). De kleurgradiënt geeft het aandeel loof- of naaldhout in een lot aan: 100% loofhoutloten zijn rood, 100% naaldhoutloten zijn blauw.

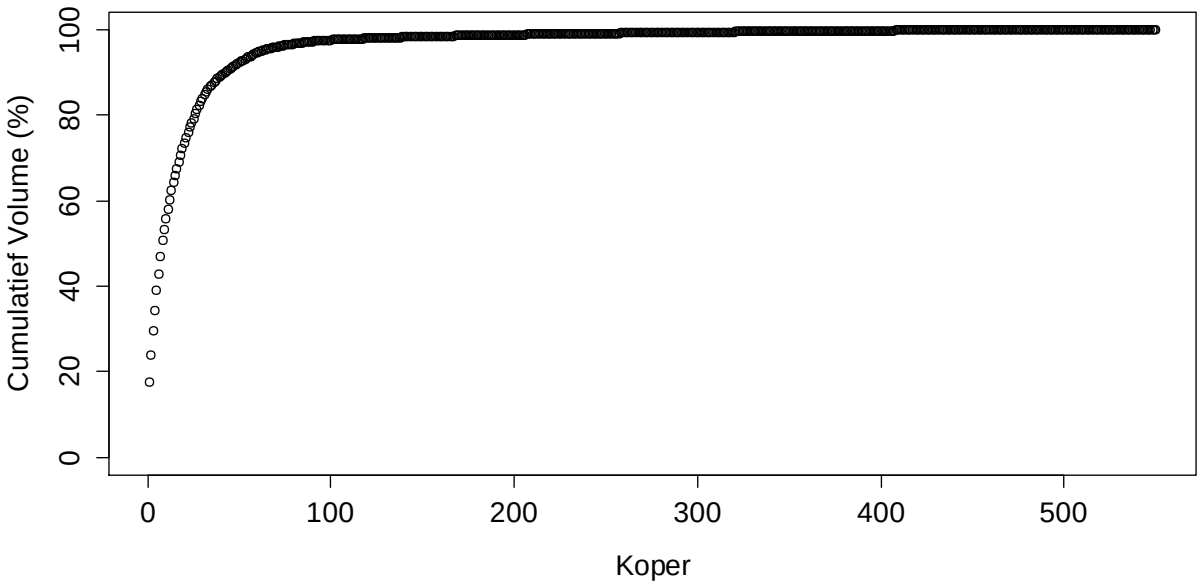


Figuur 5: Prijs (prijsniveau 2008) in functie van regio, jaar en de 12 grootste kopers (random geordend; 571 kopers in totaal)



Figuur 6: Prijs (prijsniveau 2008) in functie van FSC en Type Kap. Er is geen causaal verband tussen deze variabelen en de houtprijs

Bekijken we ten slotte nog het cumulatief totaal verkochte volume in functie van koper – gerangschikt van groot naar klein gekocht volume, zien we dat de acht grootste kopers 50% van het volume kochten. Bij 62 van de 571 kopers (10.8 %) zitten we al aan 95% van het volume.



Figuur 7: Cumulatief verkocht houtvolume (2008-1012) in functie van het aantal kopers, gerangschikt volgens dalend afgenomen volume.

3.3 Opbouw van het model

3.3.1 Conceptueel model

We kozen voor een 'mixed model', waarin naast vaste variabelen ook random variabelen voorkomen. Het vaste deel bestaat uit een sommatie van lotkenmerken (samenstelling en volume). Het random deel modelleert de structuur van de residuelen: loten verkocht binnen dezelfde regio, hetzelfde jaar of door eenzelfde koper zijn immers geen onafhankelijke waarnemingen. Het meest uitgebreide model ziet er als volgt uit:

$$\begin{aligned} Prijs(€/m^3) \sim & -1 \\ & + vf_{Beuk1} + vf_{Beuk2} + \cdots \\ & + vf_{Fijnspar1} + vf_{Fijnspar2} + \cdots \\ & + Volume{:}\,LoofNaald \\ & + (1|Regio) + (1|Jaar) + (1|Koper) + (1|FSC) + (1|TypeKap) \end{aligned}$$

met *Prijs(€/m³)* de prijs per m³ van het lot, herrekend naar het referentiejaar 2008, en $\boldsymbol{vf_{i1}}$ de volumefractie van boomsoort ***i*** uit omtrekklasse **1** (zie Figuur 2). De '-1' duidt aan dat we geen algemeen intercept in het model opnamen. Hierdoor zijn de coëfficiënten voor de volumefracties van elke boomsoort-omtrekklasse combinatie te interpreteren als de geschatte prijs per m³, in een klein lot. Voor grotere loten wordt de prijs dan nog gecorrigeerd met de volumecoëfficiënten. Voor de overige termen verwijzen we naar deel 3.2.
In een eerste stap werd de randomstructuur geselecteerd. De random intercepten voor Regio, Jaar, Koper, FSC en TypeKap werden één voor één weggelaten en vergeleken met het volledige model (voor details, zie Zuur et al. 2009). De random effecten voor Regio, Jaar en Koper bleven behouden. In het apart model voor populier werden enkel random effecten voor Jaar en Koper behouden.

3.3.2 Modeloptimalisatie

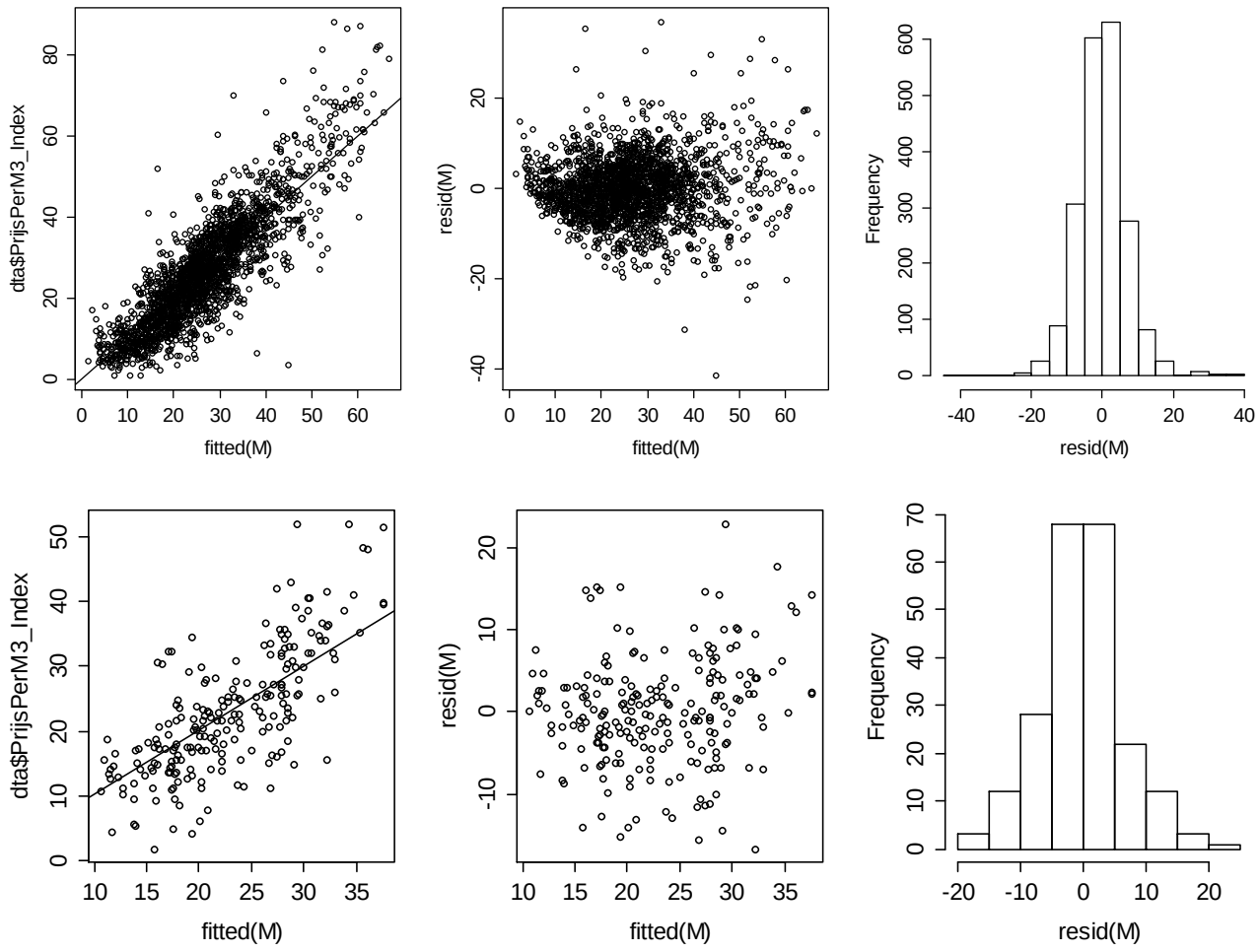
In een tweede stap werden de omtrekklassen per boomsoort vereenvoudigd door de klassen te groeperen in verschillende combinaties. Hierbij werd begonnen bij de laagste en hoogste klassen, omdat in die klassen weinig observaties beschikbaar zijn (zie Figuur 2). Verschillende mogelijke combinaties werden uitgetest. De modellen werden met elkaar vergeleken op basis van het Akaike Information Criterion (AIC), waarbij een lagere AIC een meer waarschijnlijk model aangeeft rekening houdend met het aantal gebruikte parameters. Het aldus verkregen 'optimale' model ziet eruit als volgt:

$$\begin{aligned} Prijs(€/m^3) \sim & -1 \\ & + vf_{Aloof1} + vf_{Aloof2} + vf_{Aloof34} \\ & + vf_{AmEik12} + vf_{AmEik3} + vf_{AmEik456} \\ & + vf_{Anaald12} + vf_{Anaald3} + vf_{Anaald4} + vf_{Anaald56} \\ & + vf_{Berk} \\ & + vf_{Beuk12} + vf_{Beuk34} + vf_{Beuk5} + vf_{Beuk6} + vf_{Beuk7} \\ & + vf_{CorsDen12} + vf_{CorsDen3456} \\ & + vf_{Douglas1234} + vf_{Douglas567} \\ & + vf_{Es12} + vf_{Es3} + vf_{Es45} \\ & + vf_{Esdoorn1} + vf_{Esdoorn2345} \\ & + vf_{Fijnspar12} + vf_{Fijnspar345} \\ & + vf_{Gden12} + vf_{Gden345} \\ & + vf_{InlEik12} + vf_{InlEik3} + vf_{Inleik456} \\ & + vf_{Larix12} + vf_{Larix3} + vf_{Larix4} + vf_{Larix56} \\ & + vf_{Tkastanje123} + vf_{Tkastanje45} \\ & + vf_{Boskers12} + vf_{Boskers3} + vf_{Boskers4} \\ & + Volume{:}\,LoofNaald \\ & + (1|Regio) + (1|Jaar) + (1|Koper) \end{aligned}$$

en voor populier:

$$Prijs(€/m^3) \sim -1 + vf_{populier123} + vf_{populier4567} + (1|Jaar) + (1|Koper)$$

Voor populier werd geen coëfficiënt voor volume behouden, omdat deze de onzekerheid op de prijsschattingen van de volumefracties sterk verhoogde, een probleem gekend als *multicollinearity*, dat optreedt als de variabelen in het model niet onafhankelijk zijn van elkaar. In een laatste stap werd de validiteit van het model grafisch gecontroleerd (Figuur 8). De geobserveerde waarden worden goed gemodelleerd, er is geen systematische afwijking zichtbaar in de residuelen en het histogram van de residuelen benadert een normale verdeling.

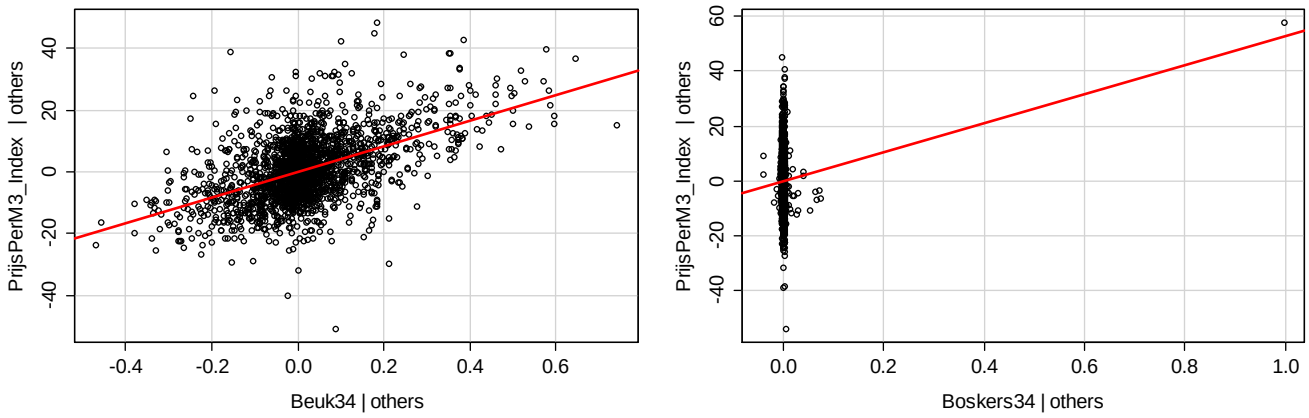


Figuur 8: Diagnostische figuren van de modelresultaten voor het model zonder populier (boven) en het model voor populier (onder): geobserveerde prijzen in functie van de gefitte prijzen (links), residuelen in functie van gefitte prijzen (midden) en een histogram van de residuelen (rechts).

Hoewel de modelselectie op basis van AIC waarden *overfitting* – het selecteren van teveel parameters – in principe tegengaan, voerden we op vraag van de stuurgroep bijkomend 10-voudige kruisvalidaties uit om de selectie van omtrekklassen te herevalueren. Gezien dit niet kan op mixed models, gebruikten we hiervoor lineaire modellen (enkel de vaste variabelen uit het model hierboven). Een kruisvalidatie beoordeelt de 'voorspelwaarde' van het model. Wanneer extra parameters leiden tot een lagere 'kruisvalidatie residuele kwadratensom', zal het opnemen van deze parameters in het model ook leiden tot correctere voorspellingen. Deze bijkomende tests beoogden voornamelijk het herbekijken van een aantal geselecteerde klassen. Zo werden de twee klassen voor Corsikaanse den bijvoorbeeld als weinig beschouwd door de stuurgroep, gezien de grote verkochte volumes (zie Figuur 2). Echter, ook op basis van deze kruisvalidaties bleken twee klassen optimaal te zijn voor Corsikaanse den. Wel suggereerde de kruisvalidaties enkele extra vereenvoudigingen in de gekozen klassen:

- Aloof1 en Aloof2 → Aloof12
- Anaald12 + Anaald3 → Anaald123
- Esdoorn1 + Esdoorn2345 → Esdoorn
- Fijnspar12 + Fijnspar345 → Fijnspar
- Larix12 + Larix3 → Larix123
- Boskers12 + Boskers3 + Boskers4 → Boskers

Op basis hiervan, en gesteund door de partiële regressieplots (Figuur 9, zie appendix voor alle plots) voerden we deze bijkomende vereenvoudigingen door. De partiële regressieplots laten toe om de robuustheid van elke geschatte coëfficiënt visueel te beoordelen. Voor elke onafhankelijke variabele Xi wordt een figuur gemaakt met op de y-as de residuelen van het model zonder Xi, en in de x-as de residuelen van een model met Xi als een functie van alle andere onafhankelijke variabelen. In deze figuren is de richtingscoëfficiënt van de regressielijn gelijk aan de geschatte prijs voor die houtsoort-omtrekklasse combinatie. De puntenwolk toont de data waarop die prijsschatting gebaseerd is. Zo zien we bijvoorbeeld voor BosKers34 dat de prijsschatting grotendeels op één lot gebaseerd is. Dit is een lot waarin alleen BosKers34 zit en daardoor een zeer grote invloed heeft op de prijsschatting van deze categorie (zie meer over invloedrijke punten in deel 3.3.3). Om die reden werd de categorie Boskers34 dan ook samengenomen met de categorie Boskers12. Een voorbeeld van een zeer degelijke prijsschatting is bijvoorbeeld Beuk34, die vertegenwoordigd is in een groot aantal loten. Ten slotte gaf de kruisvalidatie een beter resultaat met een extra volumecoëfficiënt voor naaldhout, met een opsplitsing tussen loten met 60 a 80 % naaldhout (Naald60) en loten met meer dan 80% naaldhout (Naald80).

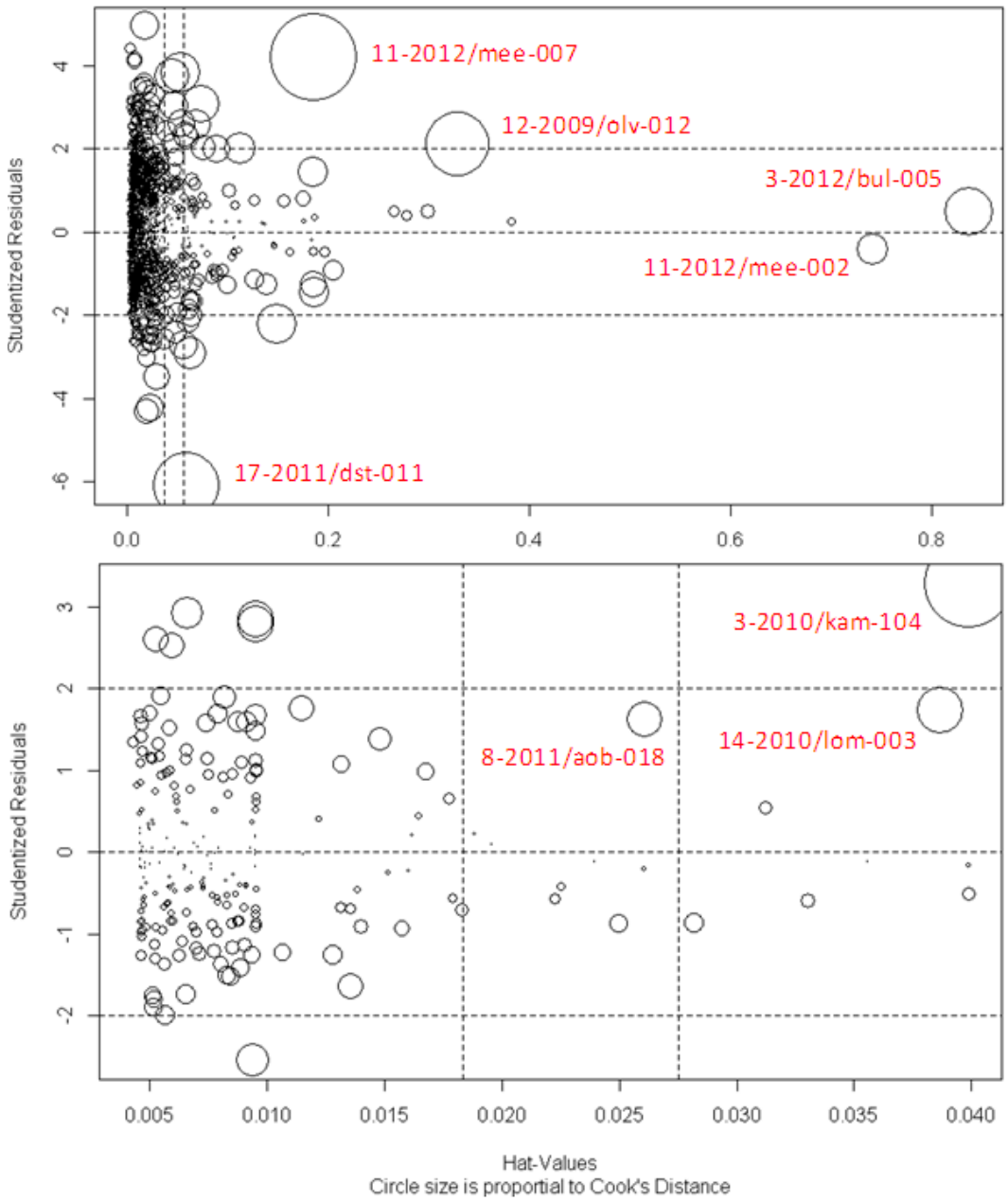


Figuur 9: Partiële regressieplots voor Beuk34 (links) en Boskers34 (rechts). De helling van de rode lijn geeft de geschatte prijs weer voor deze klassen. De data waarop deze prijsschatting gebaseerd is, is duidelijk onvoldoende voor Boskers34, waar één lot met 100% boskersen van omtrekklasse 3 en 4 zeer sterk doorweegt op de prijsschatting.

3.3.3 Invloedrijke loten

In samenspraak met de stuurgroep werd besloten om invloedrijke loten – loten die de geschatte coëfficiënten in het model sterk beïnvloeden – niet zomaar te verwijderen. Eén zeer invloedrijk lot voor populier werd weggelaten (3-2010/kam-104), omdat na navraag bleek dat dit over een lot kleine populieren ging dat al in stamstukken langs de kant van de weg was gebracht vóór verkoop.

Het is, naar de toekomst toe, sterk aan te bevelen de meest invloedrijke loten op validiteit te controleren (zie Figuur 10). Men dient zich af te vragen of de data juist werd ingegeven, er bijzondere exploitatievoorwaarden waren, de kwaliteit van het lot bijzonder goed of slecht was.



Figuur 10: Invloedrijke punten. Loten met hoge *Cook's Distance* (grootte van de cirkels), hebben een bijzonder grote invloed op de geschatte coëfficiënten. Dit zijn vaak, maar niet altijd, loten met hoge residuelen en/of *Hat values* (x-as). *Hat values* duiden outliers aan in de onafhankelijke variabelen; het zijn dus loten met een uitzonderlijke samenstelling (weinig voorkomende omtrekklassen van bepaalde boomsoorten, uitzonderlijk laag of hoog volume).

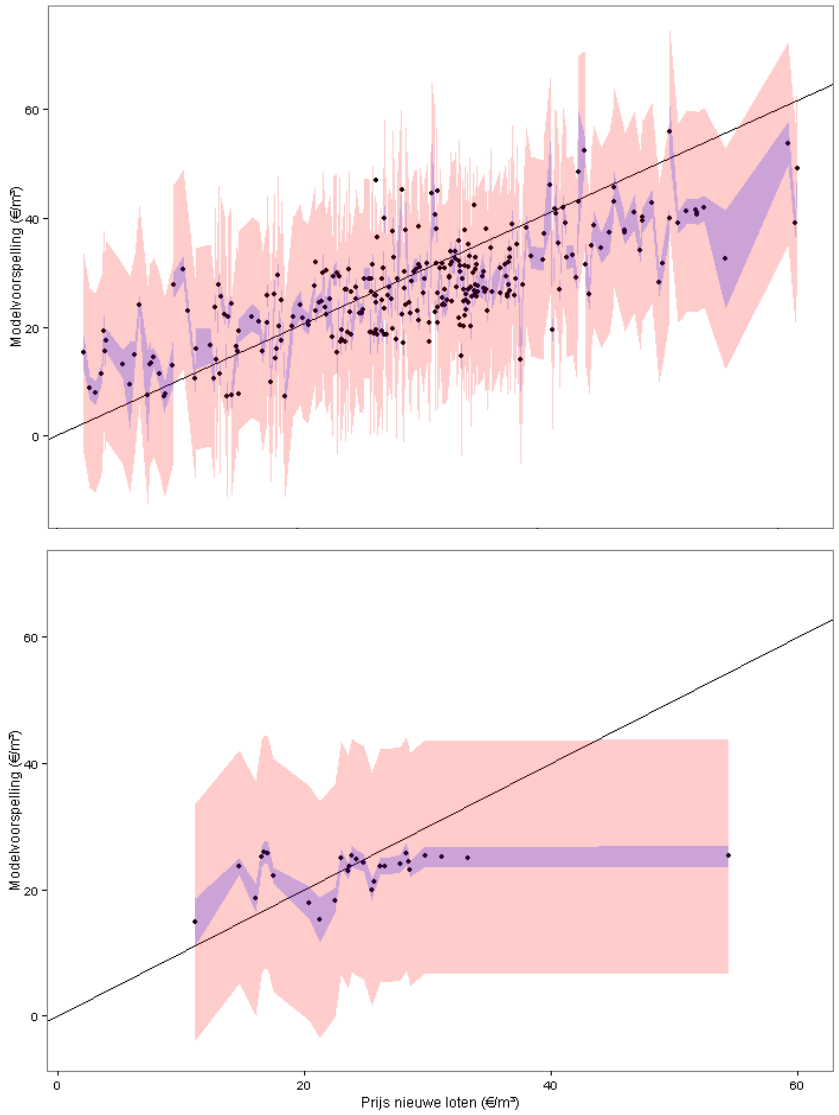
3.3.4 Validatie met nieuwe data

Met de nieuwe data van najaar 2012 en voorjaar 2013 (data2), die ANB leverde halverwege het project, werden de modelpredicties gevalideerd. Hierbij werden modelpredicties, gebaseerd op alle vaste effecten en het random effect voor Regio, vergeleken met de werkelijke verkoopsprijzen. De random effecten voor Koper en Jaar konden niet meegenomen worden, omdat in nieuwe data telkens nieuwe kopers en jaren aanwezig zijn. Bovendien wil je de prijs van een lot kunnen voorspellen onafhankelijk van wie het koopt en wanneer.

De modelpredicties over- en onderschatten de werkelijke waarden van de loten bij lage en hoge prijzen respectievelijk (Figuur 11). Dit effect is zichtbaar voor zowel het model zonder populier als voor het model voor populier, ook als we het ene extreme punt voor populier (rechts op de figuur) buiten beschouwing laten. Omdat deze data (data2) 'slechts' 333 loten omvat, is enige voorzichtigheid bij interpretatie van deze resultaten wel geboden. Het 95% gemiddeld predictieinterval is de predictie ± 18.0 €/m³ (model zonder populier) of predictie ± 18.5 €/m³ (model voor populier). Het is wellicht meer wenselijk een nauwere boven- en ondergrens te implementeren in de database, bijvoorbeeld het gemiddeld 80% predictie interval, wat overeenkomt met de predictie ± 11.9 €/m³ (model zonder populier) of predictie ± 12.2 €/m³ (model voor populier). Anders wordt de keuze van de prijszetting wel heel vrij.

Wanneer we, voor de loten gekocht door kopers reeds aanwezig in data1, het random effect voor koper bij de predicties optellen, wordt het verschil tussen predictie en werkelijkheid wel grotendeels weggewerkt. Dit wijst erop dat er nog lotkenmerken ontbreken in het model. Het kan goed zijn dat kopers die systematisch meer betalen dan verwacht op basis van de lotsamenstelling ook vaker voor kwaliteit kiezen, bijvoorbeeld voor constructiehout of fineertoepassingen. Aan het andere uiterste heb je mogelijk brandhoutkopers die weinig willen betalen, maar niets geven om de kwaliteit.

Voor de meest recente data (voorjaar 2013, 103 loten) hebben we, bij wijze van oefening, ook eens de residuen gemodelleerd in functie van de nieuwe variabelen 'Kwaliteit Type' en 'Exploitatiegraad'. Hierbij geldt 100% voor een gemiddeld lot en worden procenten bijgeteld of afgetrokken naarmate de kwaliteit of exploitatievoorwaarden beter of slechter zijn. Per 10% extra kwaliteit werden de loten 3.7 €/m³ duurder verkocht dan voorspeld op basis van de vaste variabelen alleen ($p=0.05$). Wanneer de predicties gebeuren met inbegrip van het random effect voor Regio, was het effect niet meer significant ($p=0.2$). Dit wil zeggen dat de kwaliteit verschilt tussen regio's. Exploitatiegraad had een laag positief effect, +1.3 €/m³ per 10%, maar was niet significant in beide gevallen ($p=0.5$). Deze resultaten zijn niet onmiddellijk bruikbaar, maar geven aan dat de nieuwe data voor houtkwaliteit en exploitatievoorwaarden de predicties van het model in de toekomst kunnen verbeteren.



Figuur 11: Prijzen nieuwe loten (x-as) en voorspelde prijzen (y-as) op basis van alle vaste effecten en het random effect voor Regio voor het model zonder populier (boven) en het model voor populier (onder). Punten die boven of onder de 1:1 lijn liggen zijn loten waarvoor de prijs respectievelijke over- of onderschat wordt door het model. De gekleurde zones tonen het 95% betrouwbaarheidsinterval gebaseerd op de onzekerheid van de vaste effecten (paars) en het 95% predictie interval (roze)

3.4 Resultaten

Hieronder vind je de geschatte vaste effecten (Tabel 6) in het optimale model voorgesteld in een tabel met omtrekcategoriën in cm en houtsoorten voor loof- en naaldhout. Deze zijn interpreteerbaar als houtprijs (€/m³) voor de desbetreffende categorie. De volumecoëfficiënten dienen alleen gebruikt te worden bij voorspelling van loten met <5% populier.
De random effecten voor Regio, Jaar en Koper (Tabel 7) geven aan hoeveel de prijs afwijkt van het gemiddelde. Zo is de Kust een buitenbeentje (lage verkoopsprijzen) en stegen de prijzen wat de laatste jaren (bovenop de algemene index). Koper 8 blijkt een koper die gemiddeld meer biedt voor een lot, maar het kan ook zijn dat deze koper bijvoorbeeld kwalitatief beter hout koopt (een variabele die nog niet in het model zit). Een sleutel voor de namen van kopers wordt in een apart bestand (Excel) meegestuurd.

Tabel 6: Vaste coëfficiënten (€/m³) voor de volumefracties van boomsoort-omtrekklasse combinaties en voor LN(Volume) voor Loof (>60% loof), Naald60 (>60% en <80% naald), Naald80 (>80% Naald) en Gemengd (rest van de loten)

	0-99	100-149	150-199	200-249	250-299	>300
beuk	25	37	37	43	61	33
Amerikaanse eik	24	30	41	41	41	
inlandse eik	20	26	66	66	66	
populier[1]	13	13	24	24	24	24
berk	9	9				
tamme kastanje	21	21	48	48		
es	12	19	63	63		
esdoorn	16	16	16	16		
Boskers	27	27	27			
ander loof	8	19	19			

	0-79	80-119	120-159	160-199	200-239	>240
grove den	10	16	16	16		
Corsikaanse den	8	20	20	20	20	
lariks	8	8	26	48	48	
fijnspar	15	15	15	15		
douglas	14	14	14	44	44	44
ander naald	3	3	37	39	39	

	Loof	Gemengd	Naald60	Naald80
volumecoëfficiënt[2]	0.13	0.88	1.15	1.71

[1] prijzen voor populier werden in een apart model geschat

[2] niet van toepassing op populier loten

Tabel 7: Random effecten voor Regio, Jaar en de grootste 18 kopers (€/m³) in het model zonder populier (M_a) en het model voor populier (M_p)

Provincie	Regio		Jaar			Koper		
		M_a		M_a	M_p		M_a	M_p
West-Vl.	Kust	-6.1	2008	-4.0	-2.1	1	2.6	-
	Polders - Heuvelland	-2.8	2009	-4.2	-3.7	2	-2.5	-5.0
	Brugge - Leiestreek	-0.9	2010	3.6	5.4	3	0.5	-3.6
Oost-Vl.	Oost-Vlaanderen - Noord	-2.2	2011	0.9	-1.2	4	3.8	-
	Oost-Vlaanderen - Zuid	0.3	2012	2.9	1.5	5	-4.7	-
Antwerpen	Antwerpse Kempen	-3.2				6	4.4	-
	Turnhoutse Kempen	-0.4				7	0.7	4.4
	Schelde - Neteland	-0.3				8	11.4	10.5
	Netebronnen - Zuiderkempen	1.1				9	6.5	6.1
Vlaams-Br.	Groene Gordel	0.9				10	5.9	-
	Meerdaal	0.9				11	1.3	-
	Groenendaal	2.7				12	-0.7	-
	Hageland	4.9				13	-0.7	-0.8
Limburg	Noord-Oost-Limburg	0.0				14	0.5	-
	Midden-West-Limburg	0.5				15	1.1	-
	Lage Kempen	-0.7				16	-0.9	-2.2
	Hoge Kempen	-0.1				17	4.3	-6.9
	Zuid-limburg	4.8				18	7.5	-1.0

4 Implementatie van het voorspellingsmodel

Hier geven we kort weer hoe het voorspellingsmodel kan geïmplementeerd worden voor predictie van prijzen van nieuwe loten.

4.1 Voorbereidende stappen

Boomsoorten dienen op een consistente naamgeving gecontroleerd te worden. Alle boomsoorten opgelijst in Tabel 2 worden gereclasseerd in de categorieën Aloof en Anaald. De omtrekklassen zelf hoeven niet herberekend te worden. Houd er wel rekening mee dat wanneer de 10 cm categorieën behouden blijven, er een veelvoud aan termen nodig zal zijn in de rekensom.

De volumefracties voor elke boomsoort-omtrekklassecombinatie worden berekend als het volume in het lot gedeeld door het totale lotvolume.

Om de volumecoëfficiënten te kunnen toepassen, wordt voor ieder lot ook de totale volumefractie naaldhout berekend. Dit is de som van alle volumefracties voor naaldhout over alle omtrekklassen heen. Een nieuwe factorvariabele 'LoofNaald' krijgt volgende waarden toegekend: 'Naald80', als de volumefractie naaldhout groter is dan 0.8, 'Naald60', als de volumefractie naaldhout tussen 0.8 en 0.6 ligt, 'Gemengd', met een volumefractie naaldhout tussen 0.6 en 0.4 en 'Loof', met een volumefractie naaldhout lager dan 0.4. Hierbij wordt verondersteld dat de som van de volumefracties van loof- en naaldhout in elk lot 1 is, wat het geval is als er geen data mist.

Bereken de totale volumefractie van populier in het lot (som van de volumefracties voor alle omtrekklassen)

4.2 Berekening van de predictie

Voor loten waarvan de volumefractie van populier groter is dan 0.95, wordt het model zonder populier gebruikt. Wanneer deze som kleiner is dan 0.05 wordt het model voor populier

gebruikt. In alle andere gevallen dient een combinatie gebruikt te worden en is de berekening niet volledig correct. Dit is een gevolg van de keuze voor loten met populier en loten zonder populier apart te modelleren, maar deze situatie zal slechts sporadisch voorkomen.

Berekening van de predictie voor loten met <5% populier (zie bestand 'coëfficiënten.xlsx')

- vermenigvuldig elke volumefractie met het *fixed effect* voor de overeenkomstige houtsoort-omtrekklasse combinatie en maak de som van deze producten
- vermenigvuldig LN(Lotvolume) met de juiste volumecoëfficiënt volgens de factorvariabele 'LoofNaald' en tel het bij de vorige som op
- tel het *random effect* voor de juiste regio bij de voorgaande som op

Berekening van de predictie voor loten met >95% populier (zie bestand 'coëfficiënten.xlsx')

- vermenigvuldig elke volumefractie met het *fixed effect* voor de overeenkomstige omtrekklasse van populier en maak de som van deze producten

Berekening in alle andere gevallen (zie bestand 'coëfficiënten.xlsx'):

- vermenigvuldig elke volumefractie met het *fixed effect* voor de overeenkomstige houtsoort-omtrekklasse combinatie en maak de som van deze producten. Doe dit ook voor populier.
- vermenigvuldig LN(Lotvolume) met de juiste volumecoëfficiënt volgens de factorvariabele 'LoofNaald' en tel het bij de vorige som op

Finaal moet de voorspelde prijs ($Prijs_{2008}$) nog geïndexeerd worden naar het huidige prijsniveau ($Prijs_{20xx}$):

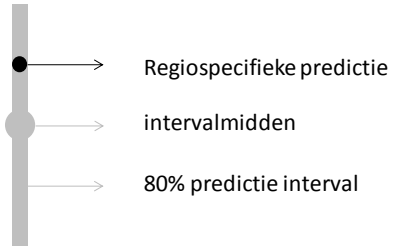
$$Prijs_{20xx} = Prijs_{2008} * \frac{Index\ 20xx_{2004}}{111.32}$$

Met $Index\ 20xx_{2004}$de huidige index met basis 2004. Zie hiervoor:
http://statbel.fgov.be/nl/statistieken/cijfers/economie/consumptieprijzen/

4.3 Berekening van het predictie interval

Behalve de predictie zelf, is het ook interessant een predictie interval te berekenen. Binnen deze grenzen kan de prijs dan wat verhoogd of verlaagd worden op basis van andere, niet gemodelleerde kenmerken als kwaliteit of exploitatievoorwaarden.
We stellen voor het 80% predictie interval te gebruiken (zie deel 3.3.4). Omdat de berekening hiervan erg complex is en er heel weinig variatie zit op de predictie intervallen, kan er gewerkt worden met de gemiddelde waarde, bekomen voor de predicties met data2.

Als intervalmidden wordt de predictie uit deel 4.2 genomen. Het maximum en het minimum worden bekomen door er 12 bij op te tellen en af te trekken. De exacte waarden (11.9 en 12.2) doen er niet echt toe aangezien zulke intervallen alleen maar een indicatie van de onzekerheid geven. Voor loten met minder dan 5% populier raden we aan de predictie zonder random effect voor regio te gebruiken als intervalmidden (zie deel 4.2). Op die manier krijgen houtverkopers een idee van de range onafhankelijk van de regio waar verkocht wordt. Bovendien moet de manuele prijscorrectie op basis van kwaliteit en exploitatievoorwaarden gebeuren op de globale predictie (zonder regio effect), omdat kwaliteit en exploitatievoorwaarden verschillen tussen regio's.



Figuur 12: Voorbeeld van een regiospecifieke predictie met een globaal predictieinterval

5 Nederlandse samenvatting

ANB verkoopt jaarlijks 200.000 m³ hout, maar er bestaat nog geen goede methodiek met betrekking tot het inschatten van de houtprijzen. Op basis van gegevens van houtverkopen tussen 2008 en 1012 wordt een prijsvoorspellingsmodel opgesteld met prijs per m³ in functie van lotkenmerken zoals samenstelling van boomsoorten en omtrekklassen, volume en regio. Dit model wordt vervolgens gevalideerd met nieuwe data van herfst 2012 – lente 2013. Ten slotte wordt een beknopt overzicht gegeven van de stappen die nodig zijn om dit model te implementeren.
De modelresultaten worden samengevat door Tabel 6, waarin de voorspelde prijzen (prijsniveau 2008) worden weergegeven per van omtrekklasse en per boomsoort. De gemiddelde prijs voor berk (steeds vrij kleine stamomtrek) is het laagst ingeschat (9 €/m³), waar die voor inlandse eik en es het hoogst ingeschat worden (66 en 63 €/m³ respectievelijk, voor een omtrek vanaf 150 cm). Het waardevolste naaldhout is lariks met een omtrek >160 cm (48€/m³). Voor loten met een belangrijk aandeel naaldhout, stijgt de lotprijs met het volume (+10 €/m³ voor een gemiddeld naaldlot van 365 m³), waar dat voor loofhoutloten niet het geval is. Er zijn ook regionale verschillen. Regio kust verkoopt 6.1 €/m³ onder de gemiddelde prijs, waar regio Hageland en Zuid-limburg 4.9 en 4.8 €/m³ boven de gemiddelde prijs verkopen voor eenzelfde lotsamenstelling.

Om, op basis van de lotsamenstelling, de geschatte prijs te berekenen, dienen de volumefracties van alle boomsoort-omtrekklassen berekend te worden en moet het totale volume gekend zijn. Ook de volumefractie van Naaldhout in het lot dient berekend to worden.

Op basis van de beperkte set loten uit het voorjaar van 2013 vonden we een significante bijdrage van houtkwaliteit – maar niet van exploitatiegraad – op de houtprijs. Deze kon niet mee opgenomen worden in het huidige model, omdat deze parameters pas sinds de invoering van de nieuwe Ivanho applicatie worden bijgehouden. Naar de toekomst toe kunnen deze parameters eveneens meegenomen worden in de update van dit model.

6 English summary

ANB sells 200.000 m³ of wood every year, but a sound methodology to estimate wood prices does not exist so far. Based on wood sale data between 2008 and 2012, a price prediction model is constructed with price per m³ as a function of lot features such as the composition of tree species and circumference classes, volume and region. Subsequently, the model is validated using new data of the autumn 2012 – spring 2013. Finally, an overview is given on how to implement the model.

7 Literatuurlijst

Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith GM (2009) Mixed Effects Models and Extensions in Ecology with R. Springer, New York