TRACK
Activating landscapes
www.track-landscapes.com

Onderzoek; representativiteit

Een van de ons meest gestelde vragen luidt: Hoe representatief is de (fiets)data? Oftewel, in hoeverre komen de intensiteiten van gebruik, resulterend uit visualisaties van appdata, overeen met de werkelijke gebruiksintensiteit?

Uiteraard bevat de appdataset waar wij gebruik van maken, slechts een klein aandeel van de (fiets) verplaatsingen in een gegeven regio. De absolute aantallen zullen daarom nooit een op een overeenkomen met de werkelijkheid. Maar dit is niet relevant, de vraag is in hoeverre activiteiten van appdata in vergelijkbare verhoudingen gebruik maken van het (fiets)netwerk. Oftewel; horen de paden die in werkelijkheid het meest intensief gebruikt worden ook in de app-dataset tot de meest gebruikte paden? En als een gegeven pad in de werkelijkheid drie keer zo intensief wordt gebruikt als een ander pad, is een vergelijkbare intensiteitsverdeling dan ook in de appdata herkenbaar?

Een manier om hier onderzoek naar te doen is de appdata te vergelijken met lokale fietstellingen, waarin werkelijke aantallen passanten gedurende een bepaalde periode geteld zijn.
Hier doet zich echter een probleem voor. In de appdata zijn recreatieve ritten oververtegenwoordigd: appgebruik ligt onder recreatieve fietsers veel hoger dan onder functionele fietsers. Deze kunnen we in de appdata van elkaar scheiden; als we een vergelijking willen maken is het nodig om functionele en recreatieve activiteiten apart van elkaar te vergelijken. Echter, bij de fietstellingen wordt doorgaans geen onderscheid gemaakt tussen functionele en recreatieve fietsers.

Daar is zover wij weten tenminste een uitzondering op: in de Provincie Vlaams-Brabant. Hier zijn (door de Directie ruimte dienst mobiliteit) op 13 locaties fietstellingen gedaan waarin ook gekenmerkt werd of het om een recreatieve of functionele fietser ging.

De studie is te vinden via: http://docplayer.nl/4722048-Eindrapport-fietstellingen-kanaalroute-zenne-provincie-vlaams-brabant-directie-ruimte-dienst-mobiliteit.html
Op de meetpunten werden (geordend van hoog naar laag daggemiddelde) onderstaande scores behaald, uitgedrukt in gemiddelde per dag.

Meetpunt	Gem_alle	recreatief	functioneel
BZ02R	413	153	260
BZ01R	363	138	225
BK07R	343	196	147
BK05R	261	131	131
BK06L	255	156	99
BK07L	181	114	67
BK02R	170	82	88
BK01L	165	64	101
BZ03R	194	99	95
BK00R	109	39	70
BK05L	173	109	64
BK06R	115	61	54
BK01R	88	26	62
SOM	2830	1366	1464

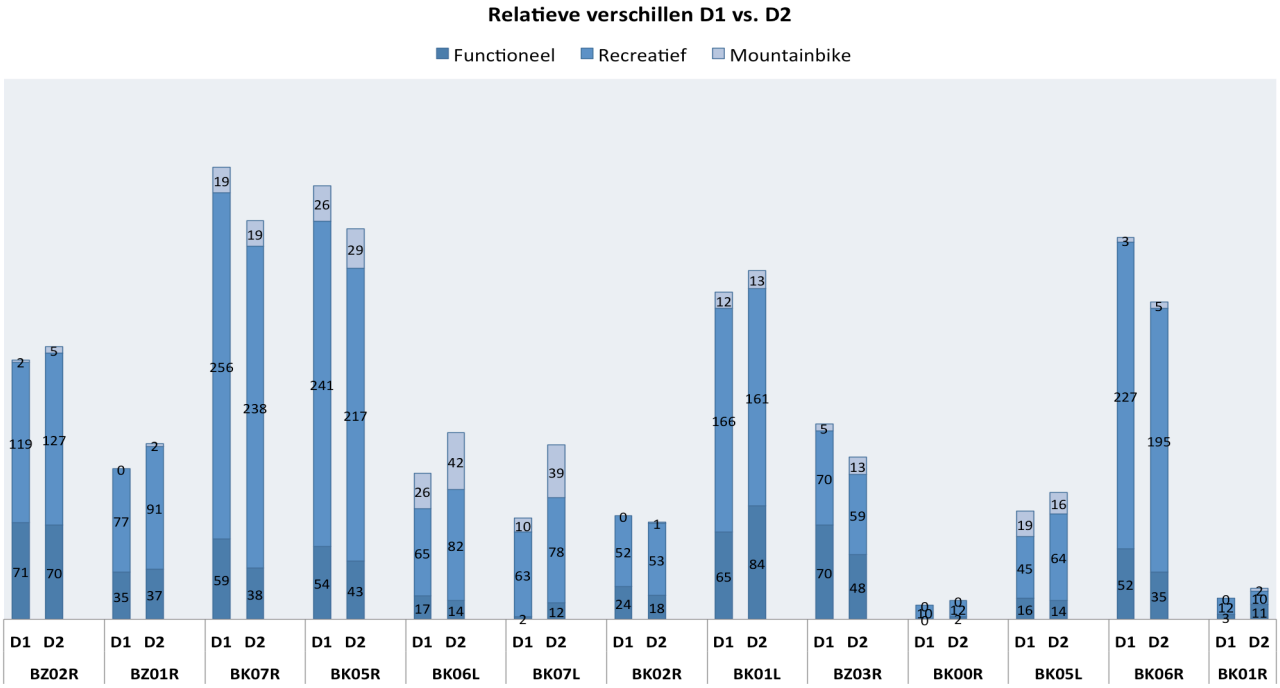
Op dezelfde telpunten hebben we uitgezocht hoeveel activiteiten er hier binnen de app-dataset passeerden.

Dataset Endomondo

Hierbij was in eerste instantie de vraag welke tijdspanne het beste genomen kan worden in de app-dataset. Het Vlaamse telonderzoek vond plaats in de maanden mei en juni van 2014. Het beste zou dan zijn om ook binnen de appdata, alleen activiteiten gedaan in mei of juni 2014 te tellen. Hier doet zich echter een mogelijk probleem voor; in een dergelijke korte periode is het aantal activiteiten binnen de appdata zeer beperkt (maximaal enkele tientallen ritten per telpunt). De andere mogelijkheid is dat we binnen de appdata, een wat grotere tijdspanne nemen; bijvoorbeeld alle data van 2014 mee te nemen. Om te weten in hoeverre dit 'erg' is, zou het goed zijn om de hoeveelheid activiteiten op de diverse meetpunten, uiteen te zetten voor twee tijdspannes: april 2014- juli2014 (evenals de Vlaamse telling, de voorjaar/zomermaanden, weliswaar aan beide kanten een maand uitgebreid) en geheel 2014.

Appdata D1 (april-juli 2014) functioneel	recreatief	mtb	D2 (heel 2014) functioneel	recreatief	mtb
41	68	1	71	127	5
20	44	0	37	91	2
34	147	11	38	238	19
31	138	15	43	217	29
10	37	15	14	82	41
1	36	6	12	78	39
14	30	0	18	53	1
37	95	7	81	161	12
40	40	3	48	59	13
0	6	0	2	12	0
9	26	11	14	64	16
30	130	2	35	195	5
2	7	0	11	10	2
269	804	71	424	1387	184

Uiteraard zijn de aantallen in D2 in absolute zin groter. Om te kunnen vergelijken of het gebruik per pad relatief hetzelfde is gebleven of is veranderd, moet daarom eerst een correctie toegepast worden op het geheel. In D1 worden daarom recreatieve activiteiten vermenigvuldigd met 0,57 (=totaal aantal activiteiten in D1/ totaal aantal activiteiten in D2=1144/1995=0,57). Daaruit volgt het volgende resultaat:



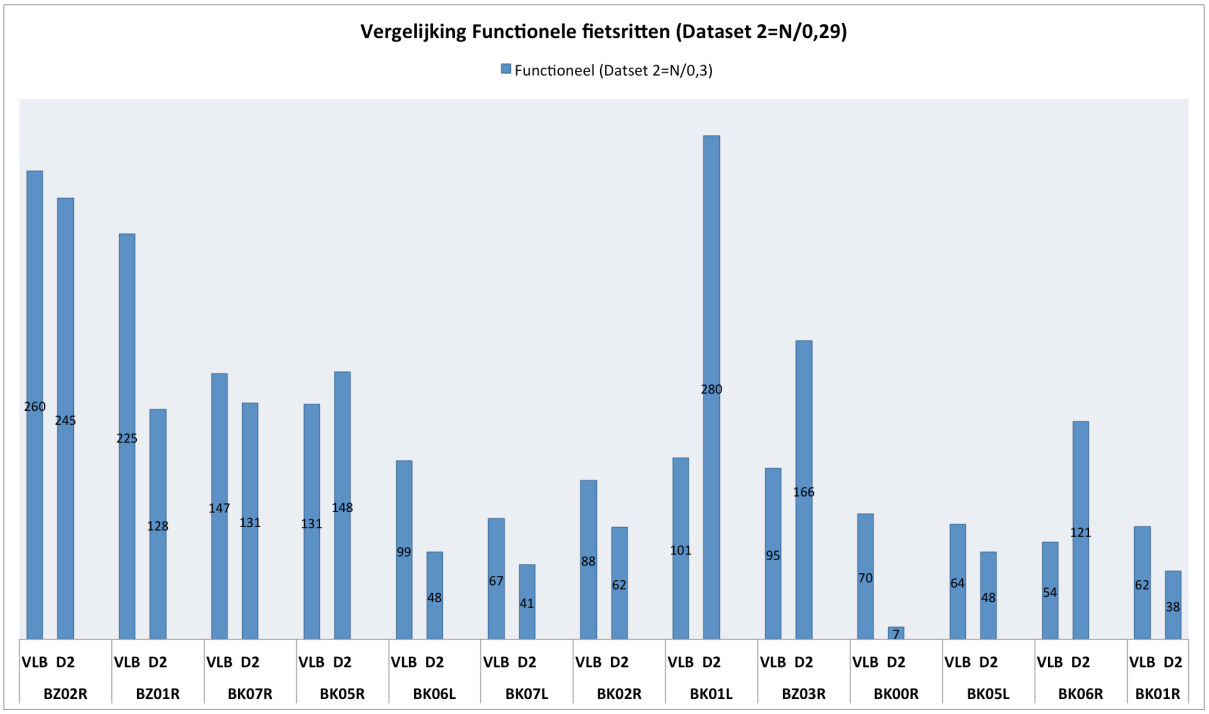
Resultaat: de verschillen zijn klein, dat betekent dat het weinig uit zou maken of we D1 of D2 zouden gebruiken voor de vergelijking met de Vlaamse telresultaten. We nemen in dit geval D2.

Vergelijking: functionele fietsritten

We beginnen met de vergelijking van de functionele fietsactiviteiten. Daarin zijn de cijfers als volgt:

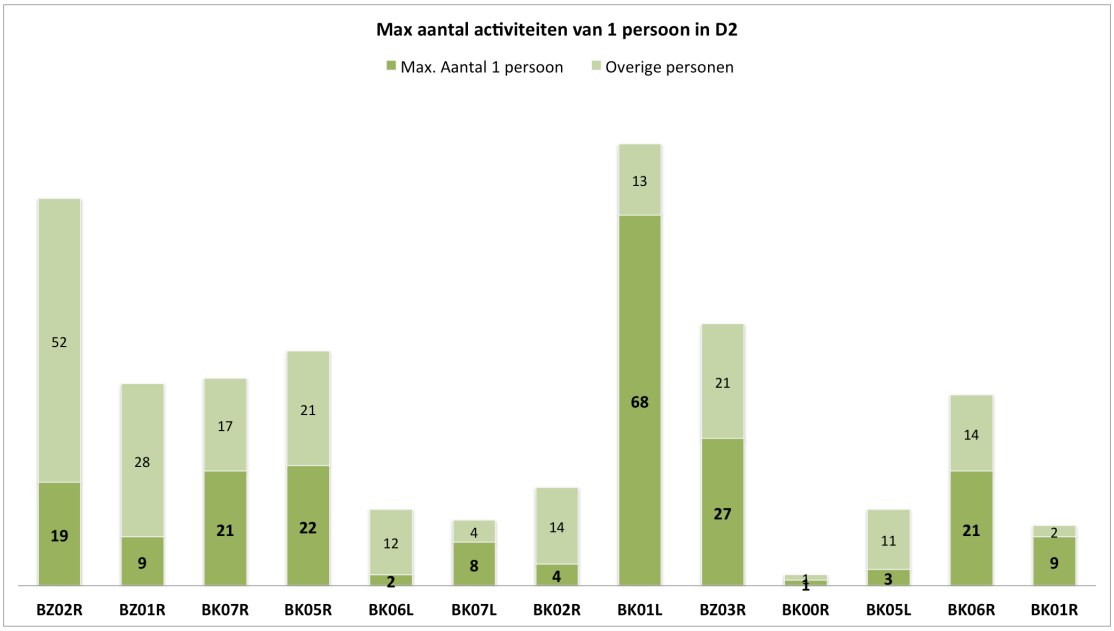
Meetpunt	VLB	D2
BZ02R	260	71
BZ01R	225	37
BK07R	147	38
BK05R	131	43
BK06L	99	14
BK07L	67	12
BK02R	88	18
BK01L	101	81
BZ03R	95	48
BK00R	70	2
BK05L	64	14
BK06R	54	35
BK01R	62	11
SOM	1464	424

De eenheden zijn hierin wel verschillend, bij de VLB tellingen gaat het om dag-gemiddelden, in D2 om totaal aantal passages in 2014. Dit is geen probleem (het gaat over een 'aantal passages per tijdseenheid'), maar om een vergelijking te kunnen maken dient wederom wel eerst een correctie toegepast te worden. De getallen van D2 worden in dit geval gedeeld door 0,29 (=totaal aantal activiteiten in D2/ totaal aantal activiteiten in VLB=1144/1995=0,29). De uitkomsten zijn dan als volgt:



Resultaat: Over het algemeen is het beeld dat de paden die binnen het VLB onderzoek intensief gebruikt werden, ook in D2 intensief gebruikt werden. Er zijn echter een aantal uitzonderingen: BK01L, BZ03R en BK06R. Hier scoort D2 significant hoger dan VLB.

In eerder onderzoek is al eens naar voren gekomen; dat een mogelijk 'gevaar' van appdata is, dat een enkel individu een grote invloed kan hebben op het algemene beeld. In D2 zijn de diverse punten tussen de 2 en 81 keer gepasseerd. Wannneer een enkel persoon dagelijks of wekelijks eenzelfde traject fietst (naar zijn werk bijvoorbeeld) en al deze activiteiten opneemt, vertegenwoordigt hij/zij al snel een groot aandeel van het aantal passages op een meetpunt. Dan kunnen vertekeningen snel ontstaan. We hebben daarom in D2 voor elk meetpunt uitgezocht, wat het maximaal aantal activiteiten van 1 persoon is:

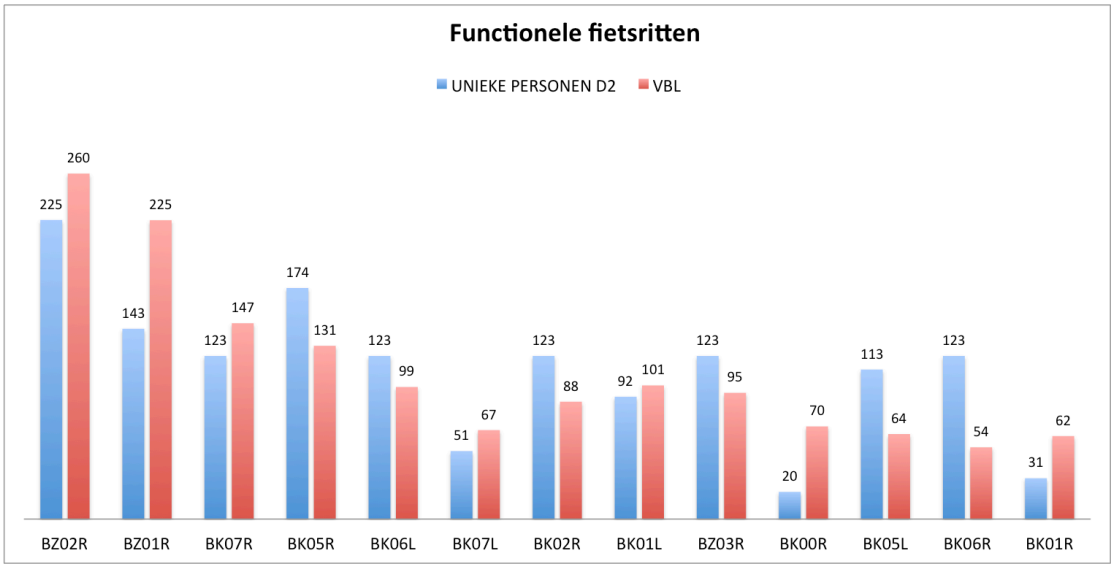


Dan wordt zichtbaar, dat bij zowel BK01L, BZ03R als BK06R, er sprake is van een hoog aantal activiteiten afkomstig van 1 individu. Bij BK01L (waarbij het verschil tussen D2 en VBL het grootst was), is ook deze verhouding het hoogste, 68 van de 81 activiteiten waren van dezelfde persoon.

Daaruit volgt de conclusie; dat de represetativiteit sterk verbeterd kan worden als er beperkingen gesteld worden aan het aantal activiteiten dat 1 persoon mag vertegenwoordigen binnen een dataset. Het is daarmee interessant om te bekijken wat de resultaten zijn, als we het aantal passages in VBL op de meetpunten, vergelijken met het aantal unieke individuen die in D2 op dezelfde meetpunten passeerden:

Meetpunt	D2: UNIEKE PERSONEN	VBL FUNCTIONEEL
BZ02R	22	260
BZ01R	14	225
BK07R	12	147
BK05R	17	131
BK06L	12	99
BK07L	5	67
BK02R	12	88
BK01L	9	101
BZ03R	12	95
BK00R	2	70
BK05L	11	64
BK06R	12	54
BK01R	3	62
	143	1464

Hier geldt wederom dat een correctie moet worden uitgevoerd, om de relatieve verdeling binnen de twee datasets over de diverse meetpunten te kunnen vergelijken. In dit geval 0,1(=totaal aantal unieke personen in D2/ totaal aantal activiteiten in VLB functioneel=143/1464=0,1). Het resultaat is dan als volgt.



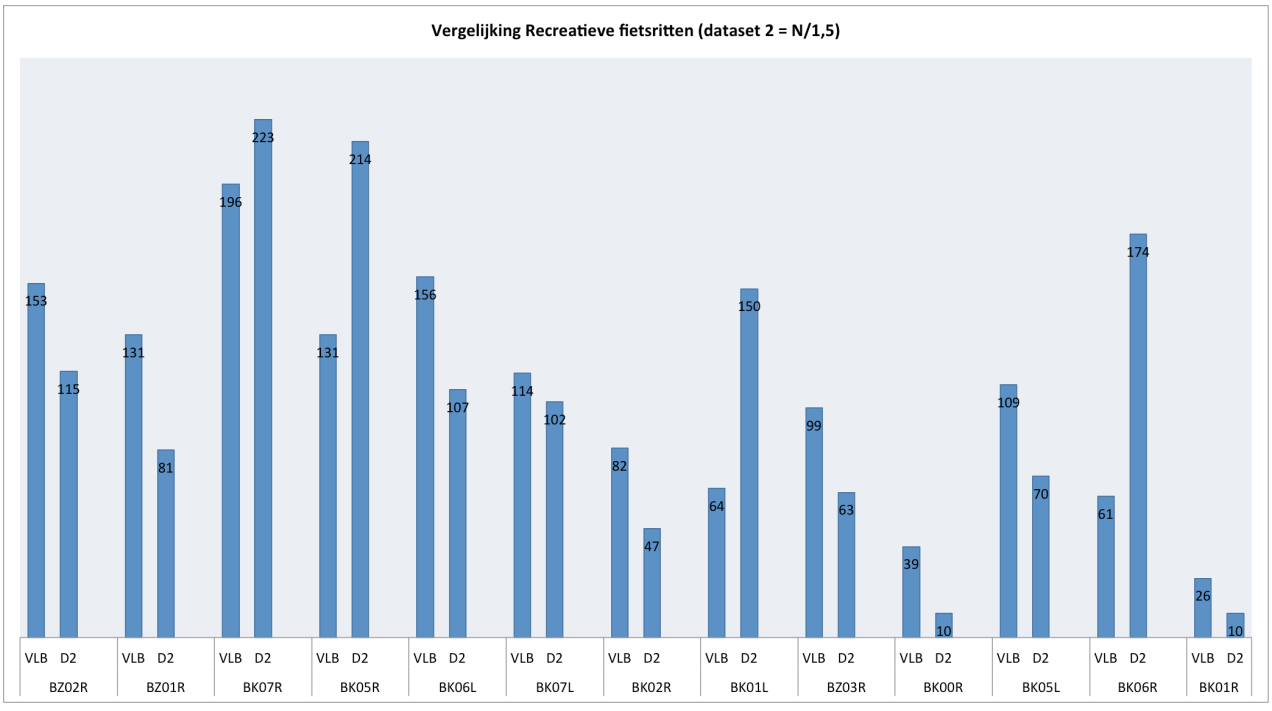
De gelijkenissen op de diverse meetpunten zijn wederom groot, maar nu zien we ook dat de uitschieters sterk gereduceerd zijn. De verschillen zijn nog het grootst in de laatste 4 meetpunten (BK00R, BK05L, BK06R, BK01R), maar hier is het aantal unieke individuen in D2 sowieso erg laag.

Vergelijking: recreatieve fietsritten

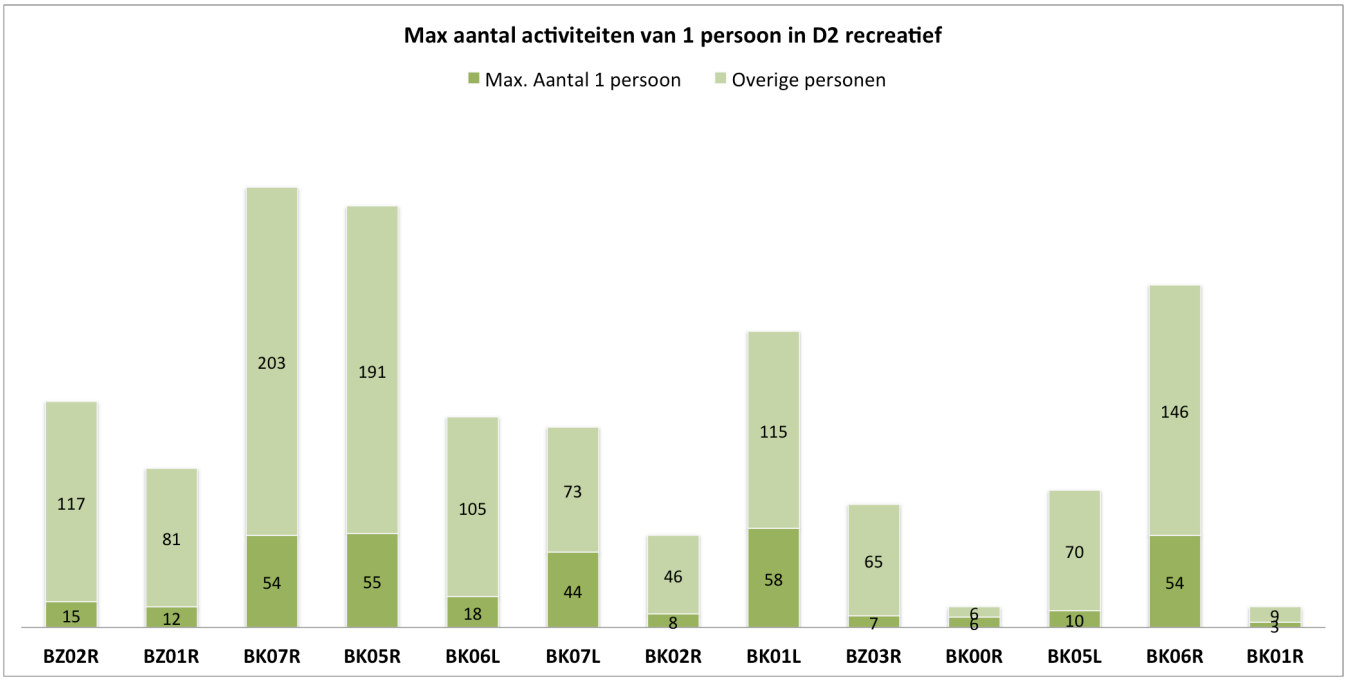
Nu volgt uiteraard de vraag, hoe zit dit bij recreatieve fietsactiviteiten? Volgens dezelfde methodiek:

Meetpunt	VBL recreatief	sum 2+3
BZ02R	153	132
BZ01R	138	93
BK07R	196	257
BK05R	131	246
BK06L	156	123
BK07L	114	117
BK02R	82	54
BK01L	64	173
BZ03R	99	72
BK00R	39	12
BK05L	109	80
BK06R	61	200
BK01R	26	12
SOM	1366	1571

Correctie is in dit geval 1,15 (=1571/1366).



Wederom zijn de meeste meetpunten behoorlijk gelijk, en zijn er enkele uitschieters.

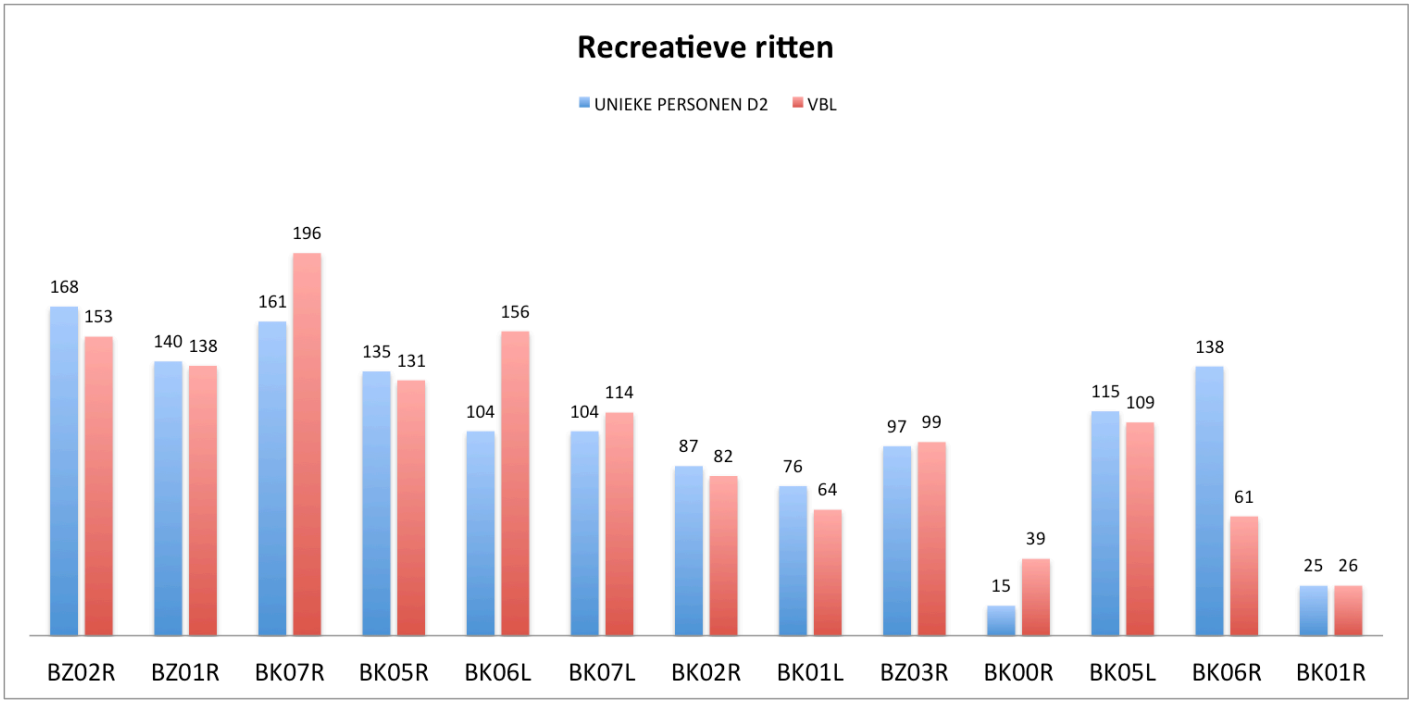


In hoeverre wordt deze vertekening bij recreatieve ritten veroorzaakt door de oververtegenwoordiging van een enkel individu?

Ook hier zien we weer eenzelfde patroon: daar waar D2 relatief meer activiteit op een meetpunt heeft dan VBL (BK05R, BK01L, BK06R), is sprake van een relatief grote vertegenwoordiging van een enkel persoon. Niet verrassend vereffent dit effect als we het aantal passages in VBL op de meetpunten, vergelijken met het aantal unieke individuen die in D2 op dezelfde meetpunten passeerden:

Meetpunt	2: UNIEKE PERSONEN	VBL
BZ02R	66	153
BZ01R	55	138
BK07R	63	196
BK05R	53	131
BK06L	41	156
BK07L	41	114
BK02R	34	82
BK01L	30	64
BZ03R	38	99
BK00R	6	39
BK05L	45	109
BK06R	54	61
BK01R	10	26
	536	1366

De correctie is in dit geval 2,55 (=1366/536).



De verschillen tussen VBL en D2 zijn op vrijwel alle meetpunten zeer klein, nog kleiner dan het geval was bij functionele fietsritten.

Hieruit volgt ook dat de verhoudingen recreatieve ritten/functionele ritten op de diverse meetpunten in D2, grote gelijkenissen heeft met VBL (wederom na het toepassen van een correctie op het geheel):

Conclusies

Vooropgesteld; de dataset die we van Endomondo in dit geval konden gebruiken voor de vergelijking is relatief klein, omdat we in België relatief weinig data hebben (minder dan in Nederland). Ook was het niet mogelijk de exact zelfde tijdspanne te gebruiken, zodoende zouden de uitkomsten nooit exact gelijk worden. Waar het hier om ging was uitzoeken of er verband zichtbaar is: en dat is er zeker.
Bij zowel functionele als recreatieve fietsritten uit de app-dataset, waren er sterke gelijkenissen met de fietstellingen uit het Vlaamse onderzoek. Echter, er ontstaan (sterke) vertekeningen als een individu een groot aandeel vertegenwoordigt van het totaal aantal activiteiten op een meetpunt. Door dit aspect te corrigeren, worden deze uitschieters sterk teruggedrongen en ontstaan een waarheidsgetrouwer beeld. Zeker bij recreatief fietsen is de gelijkenis tussen het VBL onderzoek en de app-data zeer groot. Het is te verwachten dat dit ook te maken heeft met het grotere aantal recreatieve activiteiten binnen de app-dataset. In landen waar we meer app-data hebben (onder andere Nederland) zou ook een (nog) beter beeld kunnen gaan ontstaan van functioneel fietsverkeer.

Wat betekent dit voor het gebruik/bruikbaarheid van de data?:

- Het stellen van een maximum aan het aantal activiteiten dat een individu binnen een dataset kan hebben vergroot de waarheidsgetrouwheid, aan te raden dus.
- Hou er rekening mee dat er afwijkingen kunnen zijn van enkele tientallen procenten. Dus; baseer geen beslissingen op een gedragspatroon waarin 10% verschil zit. Bijvoorbeeld; investeren in fietspad A ipv B, omdat A in de data 10% meer gebruik toont dan fietspad B, is niet sterk. Echter, als fietspad A in de appdata 50% meer gebruikt wordt dan B, is het niet waarschijnlijk dat ze in de werkelijkheid evenveel gebruikt worden. Bij een dergelijke ordegrootte van verschillen kan het een sterkere basis vormen voor een beslissing.
- Datasets van dit formaat lenen zich niet zo zeer om kleine, korte termijn trends in gebruik te monitoren, zoals dagelijkse, wekelijkse of maandelijkse veranderingen in gebruik. Echter wel om jaargemiddeldes te visualiseren, of trends door jaren heen.

Kanttekeningen:

- Dit onderzoek betreft een bepaalde regio in België. Het is niet perse gezegd dat het door te trekken is naar andere landen.
- Daarnaast; de meetpunten zijn gelegen tussen dorpskernen en bijvoorbeeld niet binnenstedelijk. Hier komen naar verwachting daardoor meer relatief lange fietsritten. We weten dat het aandeel appgebruikers ook groter is onder langere fietsritten. Dit VBL onderzoek past daarom in schaal mogelijk goed bij onze datasets. Het zou goed kunnen dat de verhoudingen binnenstedelijk wel anders liggen. Hier vormen korte fietsritten (naar verwachting) een groot aandeel van het fietsverkeer, en die zijn in onze datasets minder sterk vertegenwoordigt.