To: (10)(2e) [(10)(2e)@rivm.nl]
From: (10)(2e)
Sent: Fri 6/26/2020 7:52:16 AM
Subject: FW: Statistiek
Received: Fri 6/26/2020 7:52:17 AM

From: (10)(2e) <(10)(2e)@rivm.nl>
Sent: donderdag 25 juni 2020 15:10
To: (10)(2e) <(10)(2e)@rivm.nl>
Subject: RE: Statistiek

Hi (10)(2e)

Dank je wel – ik denk dat ik het probleem begrijp.

Ik zou een predictie analyse op basis van random forest doen. In zo'n analyse ga je de respons, nl. ziekenhuisopnames, op een dag voorspellen op basis van "achtergrond variabelen" zoals geografische coördinaten (x en y-coördinaten van de centroïde of van de grootse stad i.p.v. gemeente of departement, omdat deze waarschijnlijk te veel niveaus hebben voor random forest), aantal bewoners, aantal ziekenhuizen, bevolkingsdichtheid, leeftijdsopbouw, en liever nog anderen zoals regionale economische indicatoren, en "dynamische variabelen", zoals de temperatuur, luchtvochtigheid en luchtkwaliteit tijdens de dagen daarvóór (zeg drie dagen of een week vóór de dag waarvoor je wilt voorspellen) – of, i.p.v. deze, de gemiddeld en een schatting van algemene trend (b.v. door de Theil-Sen schatter) van deze tijdreeksen tijdens de drie of zeven dagen daarvóór, bijvoorbeeld.
Behalve deze twee groepen variabelen zou ik nog een "horloge" nemen, nl. de tijd sinds het begin van de pandemie, en "persistence forecast" variabelen, zoals de ziekenhuisopnames in de laatste drie of zeven dagen daarvóór, nemen, of de corresponderende gemiddelden en maat van trend (tijdens de laatste drie of zeven dagen vóór de dag waarvoor je de voorspelling doet). Dit is essentieel omdat het meestal verwacht wordt dat zowel de "horloge" als de "persistence forecast" de beste voorspellers zullen zijn, en de vraag is dan in hoeverre andere voorspellers nog nuttig zijn.

Als de voorspellingen goed zijn (d.w.z. als de predictie fout klein is) dan kun je naar "variable importance" kijken: welke variabelen zijn meer voorspellende? Wat ik zou verwachten is dat, behalve de "horloge" en de "persistence forecast", de economische indicatoren en de andere achtergrond variabelen meer belangrijk zijn dan de "dynamische variabelen" (temperatuur, luchtvochtigheid en luchtkwaliteit); maar zelfs dan is de vraag of deze laatste ook iets toevoegen.

Als je wilt kan ik jou helpen om een R script te schrijven om zo'n analyse te doen, maar ik zal een projectnummer moeten hebben omdat dit mij misschien drie dagen of zo gaat kosten, zelfs als je mij helpt met het voorbereiden van een data set (wat al iets eisend is, denk ik). Met de scripts kun je in principe meerdere varianten van zo'n predictie analyse proberen te doen. Wanner je interessante resultaten van de predictie krijgt kun je verder gaan met de illustratie, op basis van grafieken, van de effecten van de belangrijkste voorspellers.

Naast zo'n predictie analyse kun je in principe een associatie studie doen om associaties tussen de respons en de "dynamische variabelen" te detecteren stratificerende/corrigerende voor een aantal achtergrond variabelen. Dat zou ik ook in een script kunnen doen maar het lijkt me niet de moeite waarde te zijn: je krijgt p-waarden, maar die zeggen niet zoveel over de echte waarde van de variabelen, en het is ook iets ingewikkelder of ten minste minder algemeen omdat je de respons op een of andere manier door de tijd heen moet aggregeren.

Ik hoor graag wat je hierover vindt en of je vragen hebt.

Groeten,
(10)(2e)

From: (10)(2e) <(10)(2e)@rivm.nl>
Sent: donderdag 25 juni 2020 09:44
To: (10)(2e) <(10)(2e)@rivm.nl>
Subject: RE: Statistiek

Hoi (10)(2e)

Bedankt voor de mail, en het overnemen van (10)(2e)

Waar ik naar moet kijken is de associatie tussen transmissie van SARS-COV-2 en omgevingsfactoren (temperatuur, luchtvochtigheid, en luchtkwaliteit) voor Nederland en Frankrijk. Voor transmissie hebben we een proxy voor het reproductiegetal (Rproxy) o.b.v. dagelijkse ziekenhuisopnamen (Luo et al), voor Nederland per gemeente en voor Frankrijk per departement. De omgevingsfactoren waren rasters die ik heb geaggregeerd naar dagelijkse gemiddelden op gemeente-/departementsniveau. Verder heb ik nog een aantal demografische variabelen (leeftijdsopbouw, bevolkingsdichtheid; dit kunnen er meer worden). De analyseperiode is 16 maart t/m 10 mei (i.e. lockdowns).

We hadden bedacht om twee analyses uit te voeren: één waarin we de regio's onderling vergelijken o.b.v. gemiddelde waarden over de hele onderzoeksperiode (dus om te kijken of verschillen tussen de gemiddelde Rproxy verklaard zouden kunnen worden door een (combinatie van) verschillend klimaat of de luchtkwaliteit), en een tweede om de tijd-series te analyseren. Hierbij ligt de focus op de klimaatfactoren.

Wat ik lastig vind is het meenemen van interactie-effecten (voor de eerste analyse e.g. tussen luchtvochtigheid + bevolkingsdichtheid en luchtkwaliteit). En met het analyseren van tijd-series heb ik sowieso weinig ervaring. In ieder geval de klimaatvariabelen zijn ook (ruimtelijk) geautocorreleerd, en ik weet niet hoe ik daar mee om moet gaan. Er is veel vergelijkbaar onderzoek uitgevoerd, maar daarom vindt ik het juist lastig om te bepalen welke methoden het meest geschikt zouden zijn.. dus ik hoor graag wat jouw suggesties zijn.

Hopelijk is dit een beetje duidelijk, maar als je meer informatie nodig hebt laat maar weten.

In ieder geval bedankt!

Groet,

(10)(2e)

From: (10)(2e) <(10)(2e)@rivm.nl>
Sent: woensdag 24 juni 2020 13:39
To: (10)(2e) <(10)(2e)@rivm.nl>
Subject: Statistiek

Hallo (10)(2e)
Ik heb van (10)(2e) gehoord dat je hulp nodig hebt bij de statistische aspecten van jouw onderzoek.
Kun je mij een beschrijving maken, door middel van alle daagse taal, van de doelen van jouw onderzoek, van de data (b.v. hoeveel individuen, hoeveel metingen per individu er zijn, wat voor metingen zijn ze), en van het manier waarop de data zijn verkregen (b.v. door random sampling, door experimenten)?
Misschien kun je me ook andere relevante informatie versturen.
Als ik de benodigde informatie heb ga ik erover nadenken en dan neem ik weer contact met je op.
Groeten,
(10)(2e)