

Kokeellisen tilaston kuvaus

Vapaa-ajankalastuksen kokeellinen tilasto sisältää todennäköisyyslaskennan avulla muodostettuja arvioita vuonna 2022 vapaa-ajankalastusta harjoittaneiden asuntokuntien lukumäärästä. Kokeellisessa tilastossa julkaistaan lukumäärätietoja kalastaneista asuntokunnista otanta-asetelmassa käytettyyn ositejakoon perustuen. Uuteen laskentamenetelmään perustuva tilasto julkaistaan kertaluontoisesti kokeellisena.

Tilastoyksikkö

Tiedonkeruun kohde ja tilastointiyksikkö on asuntokunta. Asuntokunnan muodostavat samassa asuinhuoneistossa vakinaisesti asuvat henkilöt.

Tilastolliset käsitteet ja määritelmät

Asuntokunnan kalastamiseksi katsotaan se, että kuka tahansa asuntokunnan henkilö on käyttänyt mitä tahansa pyydystä ainakin kerran vuoden aikana. Henkilön katsotaan kalastaneen, vaikka hän olisi vain soutanut tai ohjannut venettä toisen kalastaessa. Vapaa-ajankalastuksella tarkoitetaan kaikkea suomalaisen asuntokuntien kalastusta (mukaan luettuna ravustus) lukuun ottamatta ammattikalastajien ja heidän asuntokuntiensa harjoittamaa kalastusta.

Tilaston perusjoukko

Tilaston perusjoukon muodostavat suomalaiset asuntokunnat ($N = 2816206$). Otanta-asetelmana käytetään ositettua otantaa. Vapaa-ajankalastustilaston otosasetelmassa suomalainen asuntokuntapopulaatio jaetaan loogisesti seitsemään ositteeseen s.e. $N = N_1 + N_2 + \dots + N_7$. Kutsu osallistua tutkimukseen lähetettiin kustakin k ositteesta poimitulle n_k ($k = 1, 2, \dots, 7$) kokoiselle otokselle.

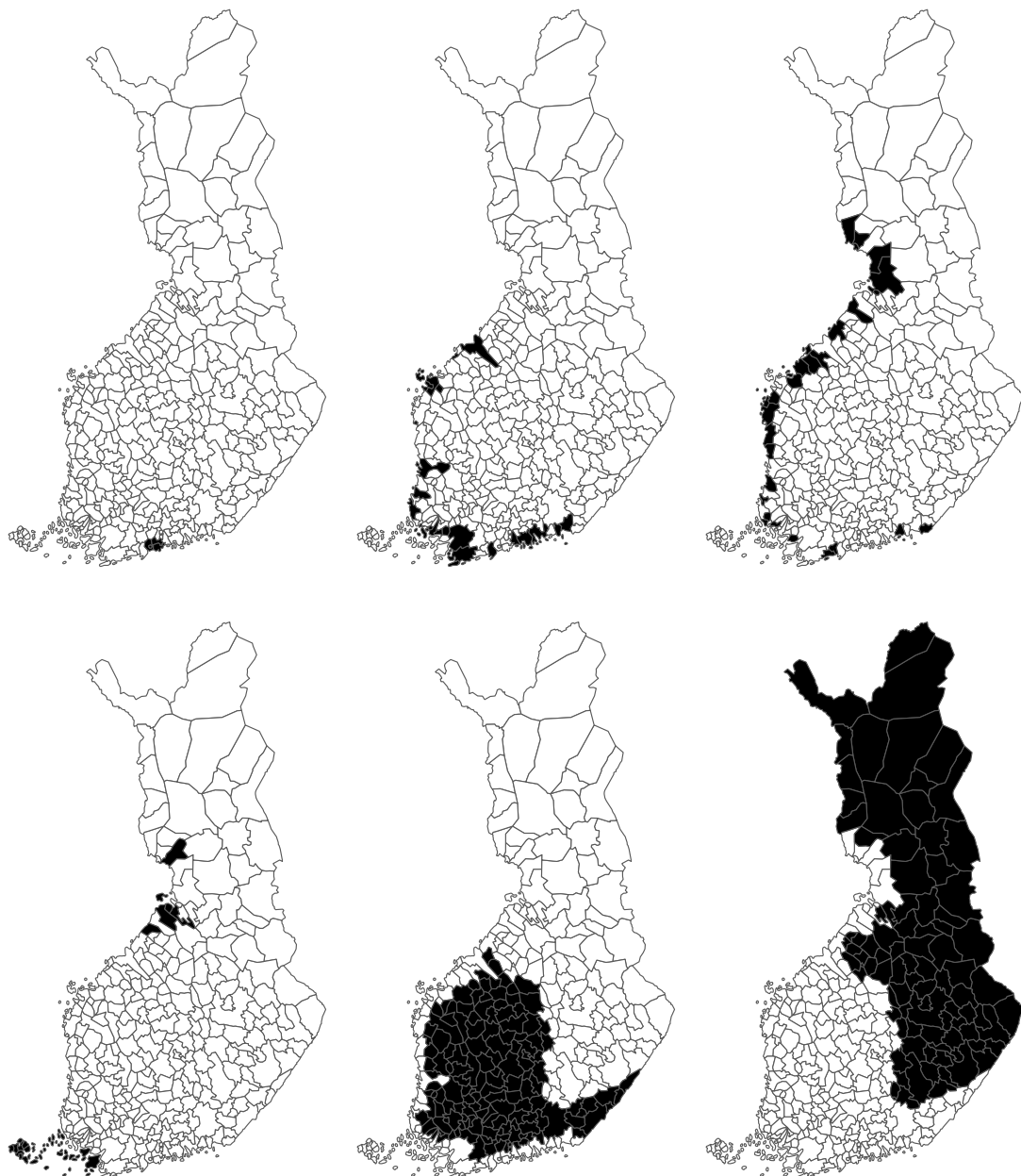
Ositteita 1-6 (kts. kuva 1) muodostettaessa hyödynnetään tietoa kohdehenkilön asuinkunnan sijainnista (pääkaupunkiseutu, muu Etelä-Suomi, Länsi-Suomi, Itä-Suomi, Oulun lääni, Lappi ja Ahvenanmaa), kuntatyyppiä (kaupunkimainen, taajaan asuttu ja maaseutumainen) sekä merellisyyttä kuvaavaa jakoa (saaristokunta, rannikkokunta ja sisämaassa sijaitseva kunta). Asuntokunnat, joihin on Metsähallituksen tietojen mukaan lunastettu kalastonhoitomaksu tilastointivuonna muodostavat oman ositteensa (osite 7). Otantaa ja arvioiden laskemista varten kalastonhoitomaksun maksaneet asuntokunnat poistetaan maantieteellisin perustein muodostetusta perusjoukkoaineistosta (ositteet 1-6). Kokonaisotos vuonna 2022 käsitti 11000 asuntokuntaa.

Tiedonkeruumenetelmä

Tiedonkeruu on kaksivaiheinen. Ensimmäisessä vaiheessa otokseen valikoituneille lähetetään kyselylomake, jonka voi palauttaa postitse tai verkossa. Kyselyyn vastaamattomia, joille löytyy puhelinnumero, tavoitellaan suppeampaan puhelimitse tehtävään kyselyyn (ns. katohaastattelu).

Menetelmät

Bayes-päätelyyn ja todennäköisyyslaskentaan perustuva kokeellinen tilasto yhdistää aiemman tiedon ja kerätyn aineiston antaen päivitettyjä todennäköisyysarvioita. Bayes-päätely nojaa Bayesin teoreemaan, joka määrittelee, kuinka todennäköisyysjakaumat päivittyvät uuden tiedon myötä. Tämä menetelmä mahdollistaa dynaamisen tilastollisen analyysin, jossa aiempi tietämys tai hypoteesit (priori) päivitetään havaitun aineiston perusteella (posteriori). Kokeellinen tilasto tehtiin käyttäen R-ohjelmistoa [3], versiota 4.2.2. Markov Chain Monte Carlo -otannat suoritettiin JAGS-ohjelmistolla [4], jota kutsuttiin `runjags` [2] R-paketin kautta. Karttakuvien piirtämisessä hyödynnettiin R-pakettia `geofi` [5].



Kuva 1: Otanta-asetelman mukaiset ositteet 1-6 järjestyksessä.

Tilastollinen malli

Otosasetelman ja tiedonkeruumenetelmän perusteella voidaan määritellä tilastollinen malli. Ensimmäisessä vaiheessa mallinnetaan otos, jonka koko $n = n_1 + \dots + n_7$ on ositekohtaisten otoskokojen summa. Merkitään asuntokunnan kalastamista suurella $K \in \{1, 2\}$ siten, että 1 tarkoittaa kalastamista ja 2 ei-kalastamista. Määritellään q_k todennäköisyytenä, että otokseen n_k valikoitunut asuntokunta ik kalastaa vähintään keran vuoden aikana. Toisin sanoen, $P(K_{ik} = 1) = q_k$ ja $P(K_{ik} = 2) = 1 - q_k$. Näin ollen suureen K voidaan ajatella noudattavan Bernoulli-jakaumaa, joka voidaan vaihtoehtoisesti ilmaista kategorisen jakauman avulla. Siten mallin kalastuskomponentti voidaan kirjoittaa seuraavasti.

$$K_{ik} \stackrel{\text{ind}}{\sim} \text{Cat}(Q_k) \quad k = 1, \dots, 7, \quad i = 1, \dots, n_k$$

$$q_k \sim p(),$$

missä $Q_k = [q_k, 1 - q_k]$ ja $p()$ on jokin jakauma. Jos asuntokunta kalastaa, niin ajatellaan että tällöin asuntokunta vastaa postikyselyyn jollain todennäköisyydellä p_{11} ja vastaavasti jos ei kalasta, niin vastaa postikyselyyn jollain todennäköisyydellä p_{12} . Edelleen, jos asuntokunta ei vastaa postikyselyyn, niin otokseen valikoituneen asuntokunnan päämiehelle yritetään löytää puhelinnumero. Ajatellaan, että numero löytyy jollain todennäköisyydellä r . Jos numero löytyy, niin ajatellaan että kalastavan asuntokunnan päämies antaa puhelinhaastattelun jollain todennäköisyydellä p_{21} , ja vastaavasti jos ei kalasta, niin jollain todennäköisyydellä p_{22} . Kaksivaiheisen tiedonkeruun lopputulosta merkitään suurella $V \in \{1, 2, 3, 4\}$, jonka realisaatio tunnetaan jokaiselle otokseen valikoituneelle asuntokunnalle.

Tarkemmin ottaen, alkuperäiseen kyselyyn osallistumiseen ($V = 1$) ajatellaan liittyvän todennäköisyys, jonka uskotaan olevan erilainen kalastaville ja ei-kalastaville asuntokunnille. Siksi olkoon $P(V_{ik} = 1|K_{ik} = 1) = p_{11}$ ja $P(V_{ik} = 1|K_{ik} = 2) = p_{12}$, kuten edellä mainittiin. Lisäksi, jos suurella r merkitään toisen otoksen (katohaastattelut) valintatodennäköisyyttä, niin $P(V_{ik} = 2|K_{ik} = 1) = (1 - r)(1 - p_{11})$ ja $P(V_{ik} = 2|K_{ik} = 2) = (1 - r)(1 - p_{12})$, mikä vastaa tilannetta, jossa asuntokunta ei vastannut alkuperäiseen kyselyyn eikä puhelinnumeroa löytynyt. Jos puhelinnumero on saatavilla, pyritään asuntokunta tavoittamaan puhelinhaastattelua varten. Koska tiedonkeruumenetelmä muuttuu, vastaustodennäköisyyden odotetaan olevan erilainen. Todennäköisyydet vastata puhelinkyselyyn, olettaen että asuntokunta ei alun perin vastannut ja puhelinnumero oli saatavilla, ovat $P(V_{ik} = 3|K_{ik} = 1) = p_{21}(1 - p_{11})r$ ja $P(V_{ik} = 3|K_{ik} = 2) = p_{22}(1 - p_{12})r$. Neljäs eli viimeinen luokka koostuu asuntokunnista, jotka eivät vastanneet alkuperäiseen kyselyyn, joille oli saatavilla puhelinnumero, mutta jotka eivät vastanneet katokyselyyn. Neljännen luokan todennäköisyydet ovat $P(V_{ik} = 4|K_{ik} = 1) = (1 - p_{21})(1 - p_{11})r$ ja $P(V_{ik} = 4|K_{ik} = 2) = (1 - p_{22})(1 - p_{12})r$. Tämä tarkoittaa, että V_{ik} voidaan käsittää satunnaismuuttujana, joka noudattaa kategorista jakaumaa ehdollistettuna kalastamiselle K_{ik} . Määritellään matriisi P , joka sisältää edellä määritellyt luokkatodennäköisyydet.

$$P = \begin{bmatrix} p_{11} & (1 - r)(1 - p_{11}) & p_{21}(1 - p_{11})r & (1 - p_{21})(1 - p_{11})r \\ p_{12} & (1 - r)(1 - p_{12}) & p_{22}(1 - p_{12})r & (1 - p_{22})(1 - p_{12})r \end{bmatrix}.$$

Edellä kirjoitetun perusteella satunnaismuuttujalle V voidaan määritellä tilastollinen malli.

$$V_{ik} \stackrel{\text{ind}}{\sim} \text{Cat}(P[K_{ik},]) \quad k = 1, \dots, 7, \quad i = 1, \dots, n_k$$

$$p_{jl} \sim p(), \quad j, l = 1, 2,$$

missä $p()$ on jokin jakauma ja $P[K_{ik} \in \{1, 2\},]$ on matriisin P rivi, joka määräytyy asuntokunnan kalastamisen (arvioidun tai ilmoitetun) perusteella. Kalastavien asuntokuntien kokonaislukumäärä $K1_{\text{SAMP}}^k$ otoksessa n_k saadaan summaamalla yksittäiset kalastavat asuntokunnat yhteen.

$$K1_{\text{SAMP}}^k = \sum_{i=1}^{n_k} -(K_{ik} - 2)$$

Suurin osa ositepopulaatioon k kuuluvista asutokunnista ei päädy otokseen. Otuomallin tarkoituksena on oppia jotain kustakin ositteesta, ja hyödyntää tätä tietoa otoksen ulkopuolelle jääneiden asutokuntien joukossa $N_{\text{OUT}}^k = N_k - n_k$. Sen vuoksi että asutokunnat on valittu otokseen satunnaisesti, ne ovat vaihdettavissa ositteiden sisällä. Lisäksi, koska Bernoullin jakauma on binomijakauman erikoistapaus, voidaan otoksesta opittua todennäköisyyttä q_k hyödyntää, kun arvioidaan kalastavien asutokuntien lukumäärää otoksen ulkopuolella. Merkitään tätä suuretta $K1_{\text{OUT}}^k$, jolloin edellä kirjoitetun perusteella

$$K1_{\text{OUT}}^k \sim \text{Bin}(N_{\text{OUT}}^k, q_k).$$

Valittu strategia on mallintaa ensin otos, jonka koko on n_k , ja sitten käyttää kyselyotoksesta pääteltyä tietoa mallinnettaessa suurempi näytepopulaatio N_{OUT}^k , jota ei ole tutkittu. Näin ollen tulosten yleistys äärelliseen superpopulaatioon, jonka koko on $N = N_1 + \dots + N_7 = n_1 + N_{\text{OUT}}^1 + \dots + n_7 + N_{\text{OUT}}^7$, on vain muutaman summaoperaation päässä. Tämä viimeinen vaihe korostaa virallisten tilastojen ainutlaatuisuutta verrattuna tavanomaiseen tilastolliseen päätelyyn. Otosmallissa populaatioparametrit oli jo päätelty. Virallisissa tilastoissa tulokset halutaan kuitenkin yleensä julkaista kokonaisina luonnollisina lukuina. Viimeinen operaatio on yksinkertainen. Arvio kalastaneiden asutokuntien lukumäärästä suomalaisten asutokuntien joukossa saadaan laskemalla

$$K1_{\text{POP}} = \sum_k K1_{\text{SAMP}}^k + K1_{\text{OUT}}^k.$$

Jakaumien $p()$ valinta

Olennainen osa Bayes-päätelyä on ennakkotiedon hyödyntäminen. Ennakkotieto voi olla esimerkiksi asiantuntija-arvio, aikasempien tutkimusten tulos tai loogisen päätelyn seuraus. Ennakkotiedot syötetään systeemiin todennäköisyysjakauman muodossa. Tätä jakaumaa kutsutaan usein priorijakaumaksi. Aikaisempia vuosia koskevien tulosten [1] perusteella kalastamisen todennäköisyyksille q_k valittiin priorijakaumat.

$$q_1, q_2 \sim \text{Beta}(2, 7), \quad q_3, q_4, q_5 \sim \text{Beta}(3, 8), \quad q_6 \sim \text{Beta}(7, 8), \quad q_7 \sim \text{Beta}(10, 2).$$

Vuosia 2014-2020 koskevien arvioiden perusteella voidaan ajatella, että ositteiden 1 ja 2 asutokunnista (parametrit q_1, q_2) ehkä kaksi asutokuntaa kymmenestä kalastaa vähintään kerran vuoden aikana, mutta ei ainakaan yli puolet kalasta. Seuraavaa kolmea maantieteellistä ositetta vastaavat suuret q_3, q_4, q_5 , joiden osalta annettua ennakkotietoa voi luonnehtia sanallisesti sanomalla, että noin kolmasosa asutokunnista kalastaa vähintään kerran vuodessa, mutta hyvin luultavasti ei kuitenkaan puolet tai enemmän. Ositteen 6 osalta voidaan sanoa, että ehkäpä noin puolet asutokunnista kalastaa kerran vuodessa. Kuitenkin melko varmasti enemmän kuin kaksi tai kolme kymmenestä, mutta hyvin luultavasti ei kuitenkaan yli 70% asutokunnista. Kalastonhoitomaksun maksaneiden osalta (q_7) ajatus on, että jos kalastonhoitomaksu on lunastettu, niin hyvin luultavasti asutokunta kalastaa vähintään kerran vuoden aikana, mutta kuitenkaan ihan kaikki eivät kalasta. Tiedonkeruuseen liittyvien parametrien priorijakaumat määriteltiin niin ikään aikaisempia vuosia koskevien arvioiden perusteella.

$$p_{11} \sim \text{Beta}(5, 10), \quad p_{21} \sim \text{Beta}(2, 10), \quad p_{12} \sim \text{Beta}(2, 5), \quad p_{22} \sim \text{Beta}(2, 5), \quad r \sim \text{Beta}(2, 6).$$

Valitut jakaumat heijastelevat ajatuksia, joiden mukaan kalastavat asutokunnat osallistuvat kyselyyn yleisesti ottaen ei-kalastavia asutokuntia paremmin, todennäköisyys osallistua puhelinhaastatteluun on korkeampi kuin postitse lähetettyyn kyselyyn ja että arviolta alle puolelle otokseen valikoituneista asutokunnista, jotka eivät osallistu postikyselyyn, löytyy puhelinnumero. On syytä huomata, että annetut jakaumat eivät ole kovinkaan huipukkaita tai tiukasti rajattuja. Näinpä niitä voidaankin luonnehtia *lievästi* informatiivisiksi. Lisäksi on huomiotava, että alhaisesta vastausasteesta huolimatta erityisesti ensimmäiseen kysymykseen (kalastiko kotitalous vai ei) saatiin lukumääräisesti paljon vastauksia. Näin ollen ennakkojakaumien vaikutus lopputulokseen on huomattavasti havaittua aineistoa vähäisempi.

Kalastamisen määritelmä

Kalastamiseksi katsotaan se, että henkilö on käyttänyt mitä tahansa pyydystä ainakin kerran vuoden aikana. Henkilön katsotaan kalastaneen, vaikka hän olisi vain soutanut tai ohjannut venettä toisen kalastaessa. Asuntokunnan katsotaan kalastaneen, jos vähintään yksi asuntokunnan jäsen on kalastanut vähintään yhden kerran vuoden aikana.

Ajankohta

Vapaa-ajankalastuksen kokeellisessa tilastossa esitetään arvioita koskien vuotta 2022. Aineisto kerättiin Luonnonvarakeskuksen toimesta keväällä 2023. Mallin kehittäminen tapahtui vuoden 2023 aikana. Tilasto tuotettiin helmikuussa 2024.

Kirjallisuutta

- [1] Antti Sykkö. The Bayesian Approach to Official Statistics: Problems & Benefits, 2023.
- [2] Matthew J. Denwood. runjags: An R Package Providing Interface Utilities, Model Templates, Parallel Computing Methods and Additional Distributions for MCMC Models in JAGS. *Journal of Statistical Software*, 71(9):1–25, 2016.
- [3] R Foundation for Statistical Computing - R Core Team. R: A Language and Environment for Statistical Computing, 2021. <https://www.R-project.org/>, [September 23, 2023].
- [4] Sarah Depaoli and James P. Clifton and Patrice R. Cobb. Just Another Gibbs Sampler (JAGS): Flexible Software for MCMC Implementation. *Journal of Educational and Behavioral Statistics*, 41(6):628–649, 2016.
- [5] M. Kainu, J. Lehtomäki, J. Parkkinen, J. Miettinen, P. Kantanen, S. Vesänen, and L. Lahti. *geofi: Access Finnish Geospatial Data*, 2023. R package version 1.0.8.