# Kokeellisen tiedon ja laskennan yhdistäminen

Marko Sysi-Aho, 48500P*

## Tiivistelmä

Kokeelliset havainnot edustavat maailmaa sellaisena kuin se meille auvautuu, tietynlaista "todellisuutta". Laskennan avulla voidaan muun muassa vertailla tästä avautuvasta maailmasta luotuja teoriapohjaisia ilmiösimulaatioita itse ilmiöstä kerättyihin havaintoihin. Usein laskentatulokset ja kokeellinen tieto poikkeavat toisistaan, ja tässä kirjoituksessa tarkastellaan yleisiä syitä, joista erot voivat johtua. Tarkastelu alkaa johdatuksella mallintamisprosessiin, jota seuraa laskennan ja havaintojen välisten syiden erittely. Keskitymme sellaiseen laskennan ja kokeellisen tiedon väliseen suhteeseen, jossa laskut perustuvat johonkin ilmiöstä tehtyyn malliin.

*msysiaho@lce.hut.fi

## I. JOHDANTO

Evoluutio voidaan ajatella abstraktien olioiden muutokseksi ajan kuluessa. Muuttuva olio voi olla esimerkiksi proteiini, eläinlaji, yhdyskunta, kulttuuri, kieli tai tieteen ala, kuten bioinformatiikka. Yleisesti oletetaan, että oliolajit, jotka selviytyvät ympäristössään hyvin säilyvät ja huonosti menestyvät karsiutuvat suuremmalla todennäköisyydellä.

Ympäristön havainnointi on lajin eräs tapa parantaa selviytymismahdollisuuksiaan: tieto ympäristön oletetusta toiminnasta auttaa oliota ohjaamaan valintojansa menestyksekkäästi. Ihminen on etevä havainnoitsija tämän hetken maapallolla. Kykymme löytää säännönmukaisuuksia ympäristöstämme on muihin lajeihin verrattuna hyvä, ja voimme halutessamme suunnata tietoisuutemme yksinomaan havaintojen tekoihin, vieläpä monimutkaisemmalla tasolla, kuin uskoaksemme yksinkertaisemmat oliot, kuten virukset tai peräti muut nisäkkäät.

Lajin selviytymiskehys on yksi tapa tulkita ihmisen alati versoava kiinnostus ympäristön mekanismien selittämiseksi. Se mitä kutsumme tiedoksi, ymmärrykseksi tai selitykseksi on filosofinen kysymys, eikä tämän oleellisen asian pohdinta ole tämän kirjoitelman tarkoitus.

Sen sijaan rajoitumme enemmänkin kuvailuun, siihen kuinka kokeellinen tieto ohjaa teorian ja mallien muodostusta. Erittelemme myös syitä, joiden vuoksi laskentatulokset ja kokeelliset havainnot saattavat poiketa toisistaan. Jos teoria ja sen mukaiset mallit kuvaavat hyvin kokeellisia havaintoja, kutsutaan lopputulosta usein ymmärrykseksi tai tiedoksi. Esimerkiksi sanomme tietävämme, kuinka kaksi painovoimattomassa tyhjiössä yhteentörmäävää palloa liikkuvat tulevaisuudessa, jos tiedämme kuinka ne liikkuvat nyt. Tieto tässä merkityksessä tarkoittaa kuvailua; meidän ei tarvitse havainnoida ilmiötä alkutilan jälkeen, vaan voimme jo kuvailla sitä mitä tulee tapahtumaan.

Yleisesti malli on selitys jostakin ilmiöstä, eksakteissa tieteissä usein matemaattinen ja laskettavissa, ja sen tavoitteena on kuvata tai ennustaa ilmiön tilaa tai kehitystä. Osa malleista pohjautuu teorioihin, jotka syntyvät yleensä kokeellisten havaintojen kautta: halutaan asettaa havaitut ilmiöt yhdenmukaiseen kehykseen. Kun jokin teoria säännönmukaisuudesta on luotu, voidaan teorian oletuksiin pohjaten rakentaa malleja, joista edelleen voidaan laskea suureita, joita voidaan verrata uusiin kokeellisiin havaintoihin. Malli on välikappale havainnoista tehtyihin oletuksiin, todellisuudesta luomaamme selitykseen. Mallintamisprosessiin kuuluu olennaisena osana parametrien estimointi, sekä mallin testaus kokeellista tietoa vastaan. Menestyksekkäät mallit, kuten Newtonin liikelait, vakiinnuttavat asemansa tieteessä. Luonnonilmiöiden mallintamiseen liittyy usein epämääräisyyttä; eksaktisti pystytään kuvaamaan vain pientä osaa ilmiöistä.

Eksaktiluonteisissa tieteissä, kuten fysiikassa monet lait kuvaavat ilmiöitä hyvin tarkasti, mutta laskennalliset rajoitteet eivät mahdollista tarkan mallin soveltamista suuriin ja monimutkaisiin makroskooppisiin ilmiöihin. Esimerkkinä tarkan mallin vaatimasta laskentatehosta olkoon proteiinien funktionaalisen toiminnan päätteleminen DNA-säikeen kodonien koodaamista aminohapoista: vaikka tiedämme kunkin kodonin koodaaman aminohapon, on aminohappojen järjestyksiä suunnaton määrä ja vain hyvin pieni osa järjestyksistä tuottaa funktioltaan halutunlaisen proteiinin. Jos tutkitaan laskennallisesti, mitkä aminohappokombinaatiot kaikista mahdollisista tuottavat kokonaisen eliön tarvitsemat proteiinit, on urakka suunnaton. Tämän ongelman kohtaavat mm. pussihukan henkiinherättäjät, jotka haluavat luoda sukupuuttoon kuolleen pussihukan löytämänsä pirstoutuneen DNA-säikeen perusteella. Jo tällaisessa tapauksessa kombinaatioiden läpikäynti on laskentatehon puitteissa hyvin hidasta, puhumattakaan siitä, että tarkastelu pohjattaisiin kodoneita alemmalle tasolle.

Tulevassa tekstissä aloitamme kokeellisen tiedon ja laskennan yhdistämisen tarkastelemalla malleja, jotka rakennetaan enemmän tai vähemmän tapauskohtaisten olettamusten pohjalta. Sen jälkeen erittelemme syitä, joista kokeellisen tiedon ja laskennan väliset erot saattavat johtua. Lopuksi selitämme, kuinka kokeellisia havaintoja sinänsä voidaan käyttää laskennan pohjana, ja kuinka laskentatulokset voivat toisinaan ennustaa ilmiöiden piirteitä, jotka ovat olleet kokeellisen tiedon ulottumattomissa.

## II. MALLINTAMISESTA

Kun tarkasteltavasta ilmiöstä on oletusten pohjalta muodestettu jokin laskennallinen malli, halutaan mallin käyttökelpoisuus varmistaa tai testata vertaamalla laskentatuloksia kokeellisiin havaintoihin. Jos mallin mukaiset laskentatulokset toistuvasti poikkeavat havainnoista, on malli hylättävä. Tällaisen mallin varaan ei voi rakentaa lisääntyvää monimutkaisuutta. Jos esimerkiksi molekyylidynamiikkasimulaatiot pohjautuisivat voimalakiin muotoa $F = mx^i \left(\frac{dx}{dt}\right)^j \left(\frac{d^2x}{dt^2}\right)^k$, jossa $i, j, k$ ovat oikeat dimensiot (yhtälöryhmän $i + j + k = 1$ ja $-j - 2k = -2$ ratkaisut) sallivia parametreja, ei monimutkaisissa usean hiukkasen simulaatioissa (tapausta $i = j = 0$, $k = 1$ lukuun ottamatta) olisi mitään järkeä, kun jo lähtökohta on heikko. Siksi mallin luonnin ja testauksen osuutta eksakteissa tieteissä ei voi aliarvioida.

Joissain tapauksissa, kuten klassisen fysiikan alueella on jo sovellusympäristössään ("sopivien oletusten alueella", ei käytetä klassista mekaniikka kvanttimekaanisten ilmiöiden mittakaavassa jne.) olemassa niin hyviä malleja, että niitä ei laskusovelluksissa tarvitse muuttaa tarkempaan suuntaan, mutta näin ei ole monissa muissa ongelmissa. Esimerkiksi tarkasteltaessa kehon solupopulaatioiden dynamiikkaa, ei tarkkoja malleja pystytä rakentamaan, vaan usein tyydytään kuvaamaan riippuvuuksia differentiaaliyhtälösysteemeillä, jotka pohjautuvat "järkevinä"pidettäviin oletuksiin. Tällaisissa tapauksissa kilpailevien mallien valinnassa auttaa suuresti, jos mallin ennustamia solupopulaatioita voidaan verrata kokeellisiin havaintoihin. Useat fysiikan lait ovatkin eksakteja juuri perusteellisen mallin testauksen ansiosta. Kertoimet, kuten painovoimakiihtyvyydet, permeabiliteetit ym. on voitu arvioida tarkasti, koska mallia on rakennettu ja testattu huolellisesti.

**Esim.** Havainnoidaan, kuinka kauan $t$ rautakuulan maahanputoaminen kestää korkeudelta $h$. Havaintojen pohjalta saattaisimme päätyä kuvaamaan aikaa mallilla $t = k\sqrt{h}$, jossa $k$ on mallin parametri. Jos mallin kuvausvoimaa halutaan testata, on saatava yhteys havaintoihin, ja tässä prosessissa ajaudutaan parametrin estimointiin.

Parametrien estimointimenetelmiä on paljon, mutta yleensä ne perustuvat jonkin havaintojen ja mallin välisen etäisyysmitan minimointiin. Sovelluksissa $N$-ulotteisten vektoreiden $x_a, x_b$ välisten etäisyyksien mittaukseen käytetään monesti Minkowsk-etäisyyksiä $(\sum_{i=1}^{N}(x_a(i) - x_b(i))^q)^{1/q}$, Euklidinen $q = 2$ ja Manhattan $q = 1$ etäisyys sekä maksiminormietäisyys $q = \infty$ tärkeinä erikoistapauksina. Estimoitaessa etäisyys on usein tilastollisiin tarkoituksiin sopivilla lisukkeilla höystetty. Esimerkiksi Mahalanobis-etäisyys $(x_a - x_b)^T\Sigma^{-1}(x_a - x_b)$, jossa $\Sigma$ on satunnaismuuttujien (vektoreissa $x$) kovarianssimatriisi tai sen estimaatti, on yksinkertainen tilastollisissa tarkasteluissa käytetty etäisyysmitta. Toisinaan tarvitaan erikoisia mittoja, joita voidaan konstruoida etäisyyden ominaisuuksiin pohjaten. Esimerkiksi silloin, kun halutaan havainnollistaa korrelaatiomatriisin alkioiden avulla kappaleiden välisiä etäisyyksiä tasossa, on korrelaatiomatriisin alkiot jollain menetelmällä kuvattava etäisyydeksi. Esimerkiksi eri valkosolulajipopulaatioiden lineaarista riippuvuutta toisistaan voitaisiin mitata ajan funktiona. Kerätyistä havainnoista estimoitaisiin korrelaatiomatriisi, ja lajien riippuvuutta haluttaisiin havainnollistaa tasossa siten, että kahden lajin välinen etäisyys on sitä lyhyempi mitä voimakkaammin toisen solupopulaation koko riippuu toisesta. Etäisyyskonstruktioihin voi perehtyä matematiikan peruskursseilla tai viitteistä

[3, 4].

Koska yksittäisten havaintojen mittaukseen liittyy häiriöitä, yritetään parametrit estimoida suuresta määrästä havaintoja. Yllä mainitussa putoamisaikaesimerkissä ainoa parametri $k$ voitaisiin estimoida jo yhden havainnon perusteella, mutta havainnointiin ja mittaukseen liittyvän, mahdollisesti myös itse tarkasteltavaan ilmiöön sisältyvän satunnaisuuden vuoksi, on estimointi luotettavampaa, kun tarkastellaan suurta määrää havaintoja. Toisaalta $n$ havaintoa $t_i, h_i$ johtavat usein kaikki eri parametrin estimaattiin. Siis kuinka estimointi kannattaisi tehdä? Tähän kysymykseen ei ole yhtä vastausta, vaan estimointia ja sen järkevyyttä on harkittava tilannekohtaisesti. Estimointiteoriaa käsitellään monissa tilastollisen analyysin oppikirjoissa, menetelmiä löytyy myös internetistä.

**Esim.** Kappaleen putoamisaikaesimerkin parametri. Kuvassa 1 on kuvitteelliset havaintoparit $(h_i, t_i)$, $i = 1, \ldots, 20$ piirretty $\sqrt{h}, t$ tasoon. Kuvassa oleva histogrammi esittää parametriestimaattien $\hat{k_i} = t_i/\sqrt{h_i}$ jakauman. Mittaukseen tai ilmiöön sisältävä satunnaisuus siis johtaa eri estimaatteihin. Tässä tapauksessa on mahdollista valita varsinaiseksi parametriestimaatiksi esimerkiksi estimaattien keskiarvo tai mediaani. Voimme myös hyväksyä estimaatiksi koko jakauman.

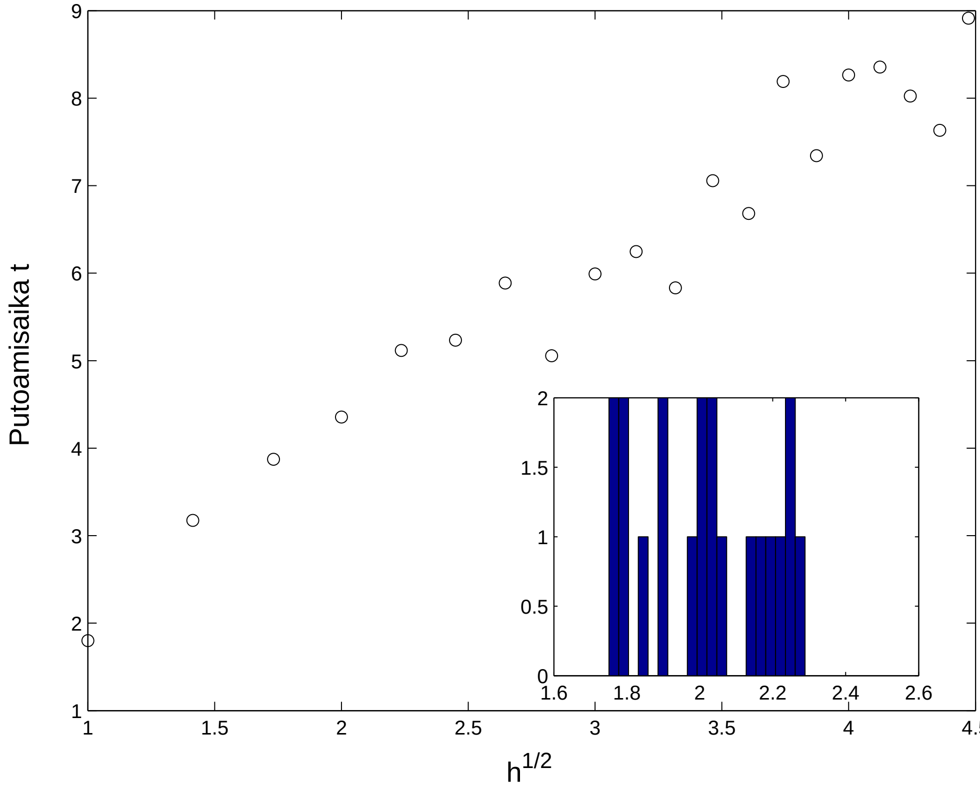
Pienimmän neliösumman menetelmä (PNS) on yleinen estimoinnissa käytetty kriteeri, jolla etsitään mallin ja havaintojen erotuksen neliösumman pienintä arvoa parametreja muuttaen. Tämä menetelmä on suosittu yksinkertaisuutensa vuoksi. Jos havainnot ovat $x_i, y_i$ ja parametreista $\Theta$ riippuva malli $f(x_i|\Theta)$, pyritään parametrit valitsemaan siten että $\sum(f(x_i|\Theta) - y_i)^2$ minimoituu. Putoamisaikaesimerkin tapauksessa estimoidaan $k$ PNS-menetelmällä minimoimalla $\sum_{i=1}^{20}(k\sqrt{h_i} - t_i)^2$ $k$:n suhteen. Tämä antaa $k = 1.98$. Havainnot oli generoitu käyttäen $k = 2$, ja lisäämällä loppuun normaalijakautuneita satunnaistermejä.

**Makrotason biologisaiheisia mallintamisesimerkkejä**

- Lääkkeen imeytyminen kehoon [1]. Annetaan henkilölle lääkeannoksia ja tarkastellaan lääkkeen konsentraatiota veressä ajan funktiona. Piirtämällä kokeelliset havainnot lin-log asteikolla, havaitaan niiden noudattavan lineaarista riippuvuutta. Kokeellinen tieto ohjaa kuvaamaan lääkkeen imeytymistä kehoon mallilla:

$$\frac{dC}{dt} = -kC \tag{1}$$

  missä $C$ on lääkkeen konsentraatio veressä. Parametri $k$ voidaan estimoida tapauskoh-



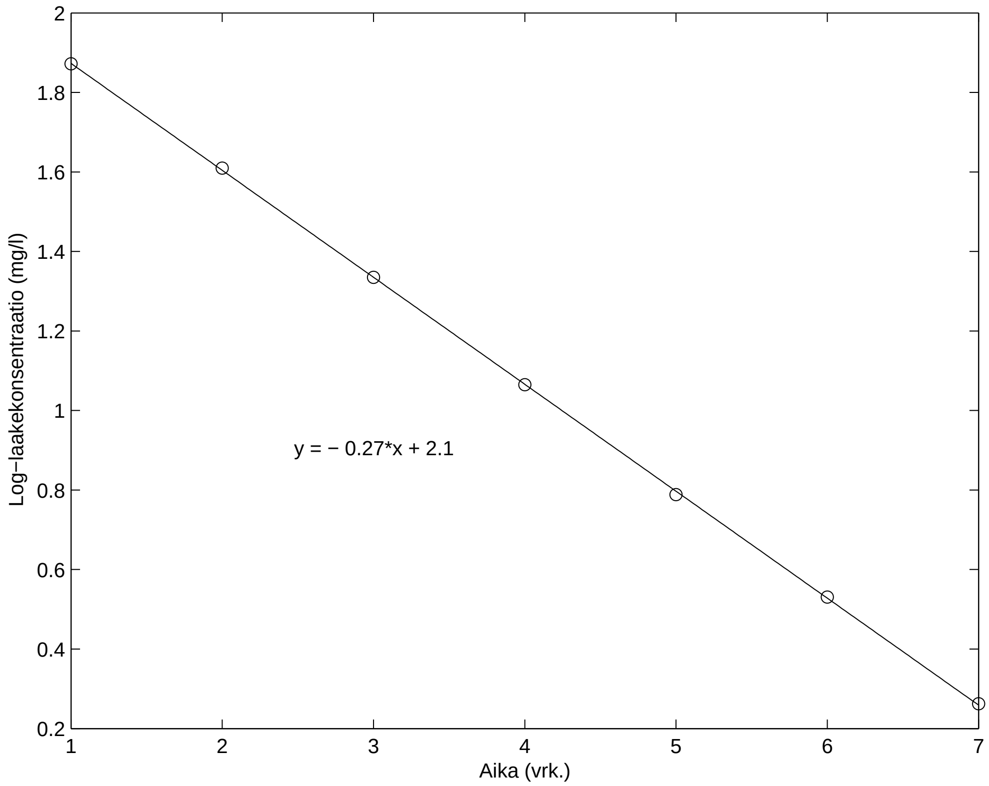

Kuva 1: Mielivaltaiset mittaustulokset kappaleen putoamisajan ja pudotuskorkeuden välisestä yhteydestä. Pienempi kuva on jakauma yksinkertaisista kulmakertoimien parametriestimaateista.

taisesti havainnoista. Malli on hyödyllinen, koska sitä voidaan soveltaa potilaan turvalliseen lääkeannosteluun: kun tiedetään, että ihmiselle turvallinen lääkekonsentraatio on välillä $[m, M]$, voidaan laskentasimulaation avulla määrätä sopivat lääkkeenottovälit ja annoskoot kunkin henkilön päivärytmiin parhaiten sopiviksi.

- Hiivapopulaation $p_n$ kasvu. Tarkastellaan hiivapopulaation kokoa ajan funktiona. Hiiva tarvitsee kasvaakseen ravintoa, mutta rajalliset ravintoresurssit eivät mahdollista populaation ylenmääräistä kasvua. Kokeellisten havaintojen rohkaisemina (Kuva 3) yritetään käyttää mallia:

$$\Delta p_n = p_{n+1} - p_n = a(M - p_n)p_n. \tag{2}$$

Mallin parametrit ovat $a$ ja $M$, joka kuvaa rajallisien resurssien olemassaoloa, kuten tässä tapauksessa ravintoa. Tämän tyyppistä lähestymistapaa sovelletaan myös moniin muihin populaatiodynamiikkamalleihin, olipa kyse sitten eläimistä, kasveista, soluista

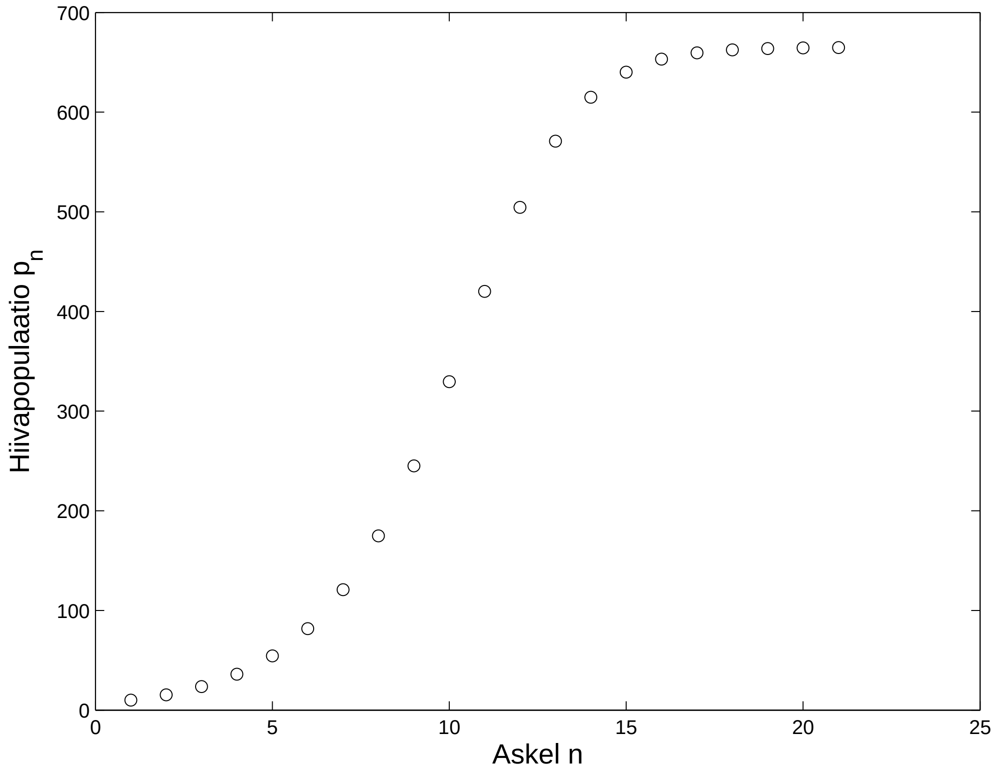



Kuva 2: Lääkekonsentraatio veressä ajan funktiona (pisteet). Yhtenäinen viiva kuvaa havaintoihin sovitettua PNS-suoraa, jonka kertoimista saadaan malliparametrit.

tai sienistä.

## III. HAVAINTOJEN JA LASKUTULOSTEN VÄLISISTÄ EROISTA

Kun teorian avulla tai ilman teoriaa luotu malli on muodostettu, voidaan sitä eksaktiluonteisissa tieteissä käyttää laskennassa ja verrata tuloksia kokeelliseen tietoon. Toivottavaa olisi, että laskennan tulokset yhtyisivät empiirisiin havaintoihin, joiden puhtaimmillaan voidaan ajatella heijastavan "todellisuutta". Havainnot siis ovat yksinkertaistaen suurin totuutemme, se mihin meidän on luotettava. Monesti laskut kuitenkin poikkeavat havainnoista. Syy voi olla käytetyn teorian tai mallin puuttellisuus, tarkan mallin yksinkertaistaminen laskentaa varten, tietokoneiden laskentatarkkuus, havaintojen keruumenetelmästä johtuvat virheet tai kaikkien näiden yhdistelmät. Tässä kappaleessa pyritään erittelemään tarkemmin kunkin syyluokan "luonnetta".

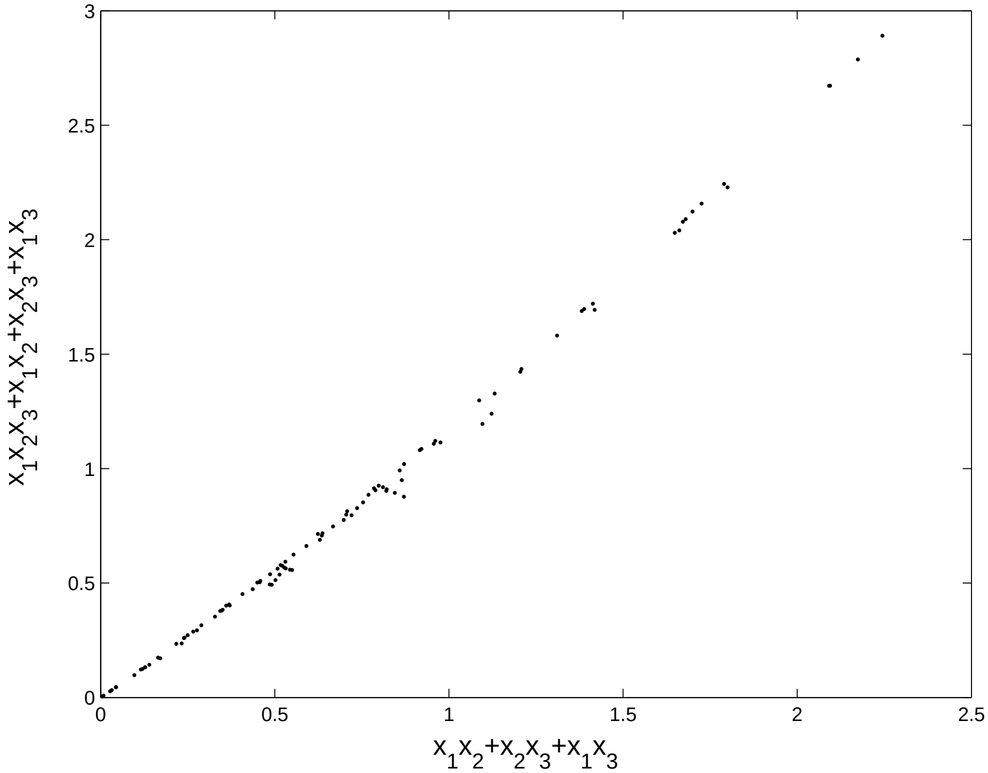



Kuva 3: Hiivapopulaation kasvu ajan funktiona, kun ravintoa on rajallisesti.

### A. Teorian tai mallin puuttellisuus

Muodostetun teorian kannalta ero laskentatulosten ja havaintojen välillä on erityisen hälyttävä silloin, kun voidaan olla varmoja, että havainnot on kerätty huolellisesti eli ne kuvaavat tarkastelumme alla olevaa ilmiötä luotettavasti ja muodostettua mallia käytetään sellaisenaan ilman yksinkertaistamista ja lisäksi tietokoneen laskentatarkkuuden vaikutus tuloksiin on tiedossa. Tällaisessa tapauksessa ero laskentatulosten ja havaintojen välillä johtuu mallin puutteellisuudesta, jonka korjaus voi vaatii teorian pohjalta tarkemmin rakennettua mallia tai teorian taustalla olevien oletusten huolellista läpikäyntiä ja mahdollisesti uuden teorian luontia.

**Esim.** Oletetaan, että järjestelmämme koostuu kolmesta vuorovaikuttavasta komponentista $x_1, x_2, x_3$, jotka voivat olla vaikkapa proteiineja, soluja tai eläinpopulaatioita. Vuorovaikutus voisi olla valkosolujen välinen kommunikointi sytokiinien välityksellä. Olemme jostain syystä kiinnostuneet näiden komponenttien vuorovaikutuksesta riippuvasta suureesta $f(x_1, x_2, x_3)$. Kuvitellaan, että räätäöimämme oletukset tai soveltamamme teoria implikoivat, että suure $f(x_1, x_2, x_3) = u_1(x_1x_2) + u_2(x_1x_3) + u_3(x_2x_3)$, jossa $u_i$ ovat joitakin funktioita,




Kuva 4: Teorian mukaisen mallin $y$-akseli ennusteet poikkeavat todellisuudesta $x$-akseli.

jotka riippuvat komponenttien pareittaisista vuorovaikutuksista. Valitaan yksinkertaisuuden vuoksi $f(x_1, x_2, x_3) = x_1x_2 + x_1x_3 + x_2x_3$. Empiiristen havaintojen perusteella kuitenkin huomataan, että malli ei kuvaa haluamaamme suuretta oikein. Havainnot poikkeavat hieman mallin ennustuksista. Olemme kuitenkin varmoja, että havainnot on tarkasti kerätty, joten niissä ei ole satunnaisvaihtelua tai virheitä. Myöhemmin aletaan epäillä, että oikea malli onkin $f(x_1, x_2, x_3) = x_1x_2 + x_1x_3 + x_2x_3 + x_1x_2x_3$ eli pareittaisten vuorovaikutusten lisäksi on myös kaikkien komponenttien samanaikaista vuorovaikutusta. Kerätään kattava otos havaintoja, ja todetaan, että näin on asian laita. Kuva 4 havainnollistaa näiden mallien eroja, kun on poimittu satunnaisesti havaintoja väliltä $x_i \in [0, 1]$. Tämän tyyppisessä tilanteessa vain pareittaisten vuorovaikutusten malliin on voitu alunperin päätyä esimerkiksi siksi, että kolmivuorovaikutustulotermi (ks. molekyylidynamiikkasimulaatioita käsittelevä osio, löydätkö tällaisen yksinkertaistuksen?) on pienehkö muihin termeihin verrattuna.

Yllä oleva esimerkki on äärimmäinen yksinkertaistus mallista, joka sisältää vuorovaikutuksia. Usein tilanne on käytännössä paljon monimutkaisempi, mutta syy laskentatulosten ja kokeellisten havaintojen välillä saattaa silloinkin olla se, että jokin vuorovaikutustermi puuttuu tai oletetaan vääränlainen vuorovaikutus. Teoriaa muodostettaessa on mahdollisesti keskitytty liaan rajoittaviin tekijöihin, kuten edellä aluksi vain pareittaisiin vuorovaikutuksiin. Saattaahan olla, että näistä vuorovaikutuksista johtuvat suureet $u_i(x_q x_p)$ ovat teorian mukaan laskettavissa, mutta todellinen järjestelmä voi sisältää myös paljon muuta kuin osiensa summan.

**Esim.** Toisena suuren mittakaavan esimerkkinä mainittakoon historiallinen uskomus siitä, että Euklidinen geometria kuvaa täydellisesti ympärillämme olevaa fyysistä avaruutta. Näin todella uskottiin pitkän aikaa. Asiaa myös epäiltiin ja ensimmäinen kokeellista todistusaineistoa keräävä tutkija oli Gauss, joka käytti kolmea vuorenhuippua ja valosignaalia näiden välillä määrittääkseen kolmion kulmien summan. (Gauss esitti aikoinaan käsitteet pintojen sisäisestä ja ulkoisesta geometriasta ja nivoi nämä yhteen Gaussin kaarevuudella, joka on tietyssä pisteessä pinnan suurimman ja pienimmän kaarevuuden tulo $\kappa = \kappa_{min}\kappa_{max}$. Euklidisen geometrian mukaan kolmion kulmien summa on $\pi$, ja ei-euklidisissa geometrioissa ylikulmaisuus $E$ kertoo kuinka paljon kulmien summa poikkeaa tästä. Kun otetaan kolmio $T$, saadaan $E(T) = \int\int_T \kappa dA$ integroimalla Gaussin kaarevuudet yli kolmion alan. Pallogeometrioissa $E > 0$ ja hyperbolisissa geometrioissa $E < 0$. Yksinkertaisimmissa, mutta silti monimuotoisissa tapauksissa Gaussin kaarevuus on vakio koko pinnan alueella.) Hän päätyi kokeessaan tulokseen, joka ei ollut Euklidisen geometrian vastainen ja päätteli, että jos fyysinen avaruus ei ole Euklidinen, se ei poikkea paljoa siitä. Myöhemmin Einstein muutti vallalla olevaa käsitystä fyysisen avaruuden geometriasta. Hänen teoriansa pohjaa massojen indusoimaan ei-euklidiseen geometriaan. Asiasta enemmän kiinnostuneet voivat katsoa viitettä [2], joka lähestyy ei-euklidista geometriaa helppotajuisesti ja tyylikkäästi kompleksianalyysin käsittein.

### B. Tarkan mallin yksinkertaistaminen laskentaa varten

Toisinaan teorian mukainen malli on voitu rakentaa hyvin tarkasti, mutta mallin käyttö laskennassa on mahdotonta toteutuksessa tarvittavien numeeristen menetelmien tai riittämättömän laskentatehon vuoksi. Tällöin sovelluksessa on voitu mallia muuttaa laskennan yksinkertaistamiseksi, ja tästä johtuen saadut tulokset saattavat poiketa kokeellisissta havainnoista.

**Esim.** Molekyylidynamiikkasimulaatioissa tarkastellaan usein järjestelmän energiaa, joka riippuu hiukkasten pareittaisista vuorovaikutuksista. Oletetaan vaikka samanmerkkiset sähkövaraukset $q_i, q_j$ kytketyiksi jousilla (lepopituus 0), ja tarkastellaan $V = \sum_{j>i} \sum_i (a\frac{q_i q_j}{r_{ij}} - br_{ij}^2)$ tyyppistä potentiaalia, jossa $r_{ij}$ on hiukkasten välinen etäisyys ja $a, b$ vakioita. Jos etsitään järjestelmän minimipotentiaalia, joudutaan yden varauksen paikkaa muutettaessa käymään läpi myös kaikki siihen kytköksissä olevat termit, ja jo $N \sim 10^5$ hiukkasella laskenta kestää kauan. Järjestelmäkoon kasvattamiseksi saatetaan tehdä laskentaa helpottavia yksinkertaistuksia, kuten tarkastella vain termejä, jotka ovat rajoittamamme alueen sisäpuolella. Esimerkkimme tapauksessa, jos vain osa hiukkasista olisi kytketty jousin toisiinsa, voitaisiin toisiinsa kytemättömien hiukkasten välisen sähköisen hylkimisvoiman ajatella olevan hyvin pieni, jos hiukkaset ovat riittävän kaukana toisistaan eli rajoittamamme alueen ulkopuolella, ja jättää tällainen vaikutus kokonaan huomioon ottamatta. Yhden hiukkasen liikkuessa tarkastelllan siten sen vuorovaikutuksia vain hiukkasiin, jotka ovat valitsemamme alueen sisäpuolella. Tällaiset malliyksinkertaistukset vaikuttavat aina havaintojen ja laskentatulosten väliseen eroon, jos alkuperäinen malli kuvaa ilmiötä tarkasti. Lisäksi edellä mainitussa esimerkissä olisi hiukkasten liikettä kuvaavat differentiaaliyhtälöt diskretoitava, ja valittava numeerinen integrointimenetelmä tulevien tilojen laskentaa varten. Kaiki tällaiset valinnat vaikuttavat enemmän tai vähemmän laskutulokseen.

### C. Tietokoneen laskentatarkkuus

Myös tietokoneen laskentatarkkuus voi olla syy havaintojen ja mallin välisille eroille. Liukuluvuilla esitetään vain äärellinen osajoukko reaaliluvuista, ja lisäksi näitä lukuja ei ole kaikkialla reaalilukuvälillä yhtä tiheästi. Liukulukuja käytettäessä suurissa laskutoimituksissa tapahtuukin suunnaton määrä pyöristyksiä, jotka yhdessä vääristävät laskentatulosta. Siten, vaikka kaikki muut laskentaan vaikuttavat tekijät olisivatkin kohdallaan, saattaa tulos poiketa kokeellisesti havaitusta yksin pyöristysvirheiden vuoksi.

Näissä tapauksissa on avuksi, jos selvittää tavan, jolla laskennassa käytetty ohjelmisto luvut esittää. Lisäksi voi miettiä, kuinka laskentajärjestyksen valitsee pyöristysvirheiden pienentämiseksi, mm. summaukset on joskus mahdollista räätälöidä siten, että termit ovat samaa suuruusluokkaa. Pyöristysvirheistä johtuvia ongelmia käsittelevät numeerisen laskennan, ohjelmointikielten ja ohjelmistojen oppaat. Apua löytää myös internetistä järkevillä hakusanoilla.

### D. Havaintojen keruumenetelmän vaikutus tarkasteltavaan ilmiöön

Koska kokeelliset havainnot muodostavat monissa tapauksissa tieteellisen tutkimuksen ytimen, olisi toivottavaa, että ne kuvaisivat luotettavasti ilmiötä, jota tarkastellaan. Hyvin kerätyt havainnot edustavat tutkijalle "todellisuutta", ja teoriat, jotka eivät selitä havaintoja, ovat käyttökelvottomia. Havaintojen keruuseen saattaa kuitenkin liittyä ongelmia. Mittaukset eivät ole absoluuttisen tarkkoja, vaan antavat mitatun suuren vain tietyllä tarkkuudella. Lisäksi mittauksiin liittyy erinäisiä virheitä tai itse mittaus voi vaikuttaa tarkastelun alla olevaan ilmiöön. Tällaisissa tapauksissa teoreettisen mallin pohjalta lasketut tulokset saattavat poiketa havainnoista myös havainnoissa olevan virheellisyyden vuoksi.

**Esim.** i) Tehdään pituusmittauksia väärin skaalatulla mittanauhalla. ii) Mittalaitteen varaus vaikuttaa sähköiseen mitattavaan suureeseen. iii) Mittalaite antaa vääriä havaintoja komponenttivian vuoksi. iv) Kerätään havaintoja typerästi, esimerkiksi tutkittaessa kemikaalin konsentraatiota tai radioaktiivisuutta kehossa otetaan näytteitä vain isovarpaasta. v) Tarkasteltavaa ilmiötä ei eristetä riittävän hyvin muista havaintosuureen arvoihin vaikuttavista tekijöistä.

## IV. HAVAINNOISTA MALLINTAMINEN

Joskus tarkastellaan ilmiöitä, joista tiedetään hyvin vähän tai ei mitään. Teorioita tai edes malleja ilmiöstä ei ole luotu, mutta kokeellista tietoa on onnistuttu keräämään. Toisinaan voidaan kokeelliset havainnot sinänsä ottaa mallin pohjaksi. Esimerkiksi konstruoidaan empiirinen jakauma, kuten histogrammi havainnoista, ja käytetään tätä jakaumaa todennäköisyysennusteiden tekoon. Tällöin laskentaa käytetään apuna kokeellisten havaintojen ennustamisessa. Tämä prosessi voi toisinaan johtaa kasvavaan tietoon, yhä tarkentuviin malleihin ja käyttökelpoisen teorian syntyyn. Esimerkiksi Kepler teki aikoinaan vain havaintoja planeettojen liikkeistä, hänellä ei ollut selvää teoriaa havaintojensa pohjalla. Hän vain totesi, että planeettojen radat ovat ellipsejä, auringosta planeettaan pirretty säde piirtää yhtä pitkissä ajoissa yhtä suuret pinta-alat ja että eri planeettojen kiertoaikojen neliöt suhtautuvat toisiinsa kuten niiden Auringosta mitattujen keskietäisyyksien kuutiot. Tämä on esimerkki havainnoista mallintamisesta. Vasta Newton myöhemmin onnistui pukemaan nämä havainnot syvällisemmiksi malleiksi ja edelleen teoriaksi.

Havainnoista mallintamisesta on kyse useissa tilastotieteen ongelmissa. Bioteknologiassakin lienee tilanteita, joissa esimerkiksi jonkun proteiinin tai genomisegmentin aktiivisuutta voidaan tutkia kokeellisesti, mutta ei tiedetä, miksi tämä segmentti tai proteiini on aktiivinen. Kuitenkin tietoa aktiivisuudesta voidaan tarvita apuna muita ilmiöitä, kuten syövän puhkeamista selitettäessä. Tai immuunijärjestelmän toimintaa tarkasteltaessa olisi hyvä tietää, millä todennäköisyydellä synnynnäisen immuunijärjestelmän komplementtijärjestelmän proteiineja esiintyy tietyn elimen ympäristössä annetulla konsentraatiolla. Tätä voidaan käyttää apuna, kun arvioidaan viruksen tai bakteerin leviämisnopeutta, sekä erinäisten valkosolujen määrää alueella. Tällaisessa tapauksessa kokeellinen tieto sinänsä auttaa ennusteiden teossa.

Geenien toiminta-aktiivisuutta tarkastellaankin jo pitäen silmällä tautien, kuten jonkun syövän puhkeamisen todennäköisyyttä. Esimerkiksi $n$ prosenttia tapauksista, joissa geeni $X$ on ollut aktiivinen on johtanut syövän puhkeamiseen. Jos geeni aktivoituu paljon ennen syövän oireita, on taudin kulku mahdollista estää paremmin.

## V. YHTEENVETO

Kokeellinen tieto ohjaa teorianmuodostusta ja tarpeellisen laskentamenetelmän ja laskentatarkkuuden valintaa. Havainnot edustavat meille "todellista"maailmaa ja monesti tieteellisen laskennan tavoitteena on selittää tätä "todellisuutta"onnistuneesti. Hyvän teorian rakentaminen on pitkä, jopa sukupolvia kestävä prosessi, johon kuuluu havaintojen keruu ja mallien muodostus. Malli on välikappale, joka yhdistää kokeelliset havainnot ja laskentatulokset. Kun tarkastellaan laskentatulosten ja havaintojen eroja, on huomio ongelma-alueesta riippuen kiinnitettävä eri tekijöihin. Molekyylidynamiikkasimulaatioissa ei keskivertotutkijan tarvitse miettiä mallin testausta tai uudelleen luontia, vaan huomio on enemmänkin kiinnitettävä jo olemassaolevan mallin järkevään käyttöön. Esimerkiksi laskentaa varten tehtävien yksinkertaistusten, kuten potentiaalikatkaisujen kanssa on oltava varovainen. Klassisen fysiikan lakeja ei kannata soveltaa pienen mittakaavan ilmiöihin, joissa kvanttimekaanisien mallien käyttö on välttämätöntä. Usein laskentaresurssit rajaavat tarkkojen mallien sovellusta. Nykytekniikalla esimerkiksi $10^{10}$ vuorovaikuttavan hiukkasen MD-simulaatiot ovat erittäin hitaita toteutettavaksi. Joskus on vain tyydyttävä tarkastelemaan riittävän pieniä järjestelmiä. Lisäksi on huomioitava tietokoneen laskentatarkkuudesta johtuvat virheet. Ne henkilöt, jotka keräävät havaintoja, joutuvat paneutumaan siihen, että se mitä kerätään edustaa ilmiötä, jota halutaan tutkia. Tilanteeseen järkevä koesuunnitelma ja luotettavat, riittävän tarkat mittaukset ovat oleellisessa osassa. Jos havainnot ovat huonosti kerätty, voi ilmiötä hyvinkin kuvaava teoria tai malli joutua kyseenalaistettavaksi vain siksi, että "todellisuutta"ei havainnoissa kuvattu oikein. Laskennan avulla voidaan lisäksi hyödyntää kokeellisia havaintoja lukuisissa tilanteissa, joissa ilmiöstä ei tiedetä juuri mitään, mutta ilmiön kehitystä voidaan kerättyihin havaintoihin pohjaten menestyksekkäästi ennustaa. Tilastotieteen lukuisat keinot edustavat tällaista lähestymistapaa. Parhaimmillaan laskentatulokset voivat ennustaa uusia ilmiön piirteitä, joita kokeellisesti ei vielä ole havaittu, mutta jotka ovat olemassa. Esimerkkeinä mainittakoon kvarkit ja aikadilaatio, jotka molemmat ennustettiin teoriassa ennen kuin niiden olemassaolosta oli kokeellista tietoa. Näin ollen kokeellisen tiedon ja laskennan välinen vuorovaikutus voi hedelmällisimmillään olla myös aidosti kaksisuuntaista.

---


[1] F. R. Giordano *A first course in mathematical modeling*

[2] Tristan Needham *Visual Complex Analysis*

[3] Ronald F. Gariepy *Modern Real Analysis*

[4] E. Kreyszig *Introductory Functional Analysis with Applications*