September 2019

SNS ANALYS nr 58

Städer och stora datamängder

NYA STORA DATAMÄNGDER möjliggör mätningar av stadskaraktäristika i både högre frekvens och bättre upplösning än någonsin tidigare. De stora datamängderna kommer dock inte att lösa stadsforskningens stora frågor på egen hand. Den största potentialen i de nya datamängderna är när de representerar det som tidigare inte kunde mätas eller när informationen kan kopplas samman med händelser som inte kunnat förutses av ekonomiska beslutsfattare (så kallad exogen variation).

AVSAKNADEN AV KOPPLINGEN mellan urbana tillämpningar av samhällsvetenskap och den faktiska fysiska staden har delvis berott på avsaknaden av mätningar av stadens fysiska attribut. Nya stora datamängder har potential att råda bot på detta.

HÄR VISAS HUR BILDER från Googles gatuvy (Google Street View) kan användas för att prediktera inkomster i olika delar av New York. Det visar potentialen i denna teknik som skulle kunna användas till att mäta exempelvis hittills okända inkomster i utvecklingsländer.

DATA FRÅN PLATTFORMAR som Yelp som laddas upp av användare ger ofta aktuellare och geografiskt bättre upplösta mätningar än officiell statistik från exempelvis Statistiska centralbyrån (SCB). Rapporten visar potentialen i hur Yelpdata kan komplettera officiell statistik i geografiskt högupplösta realtidsmätningar av olika fenomen.

FÖRFATTARE
Edward Glaeser, professor i nationalekonomi vid Harvard University.

SNS ANALYS En stor del av den forskning som bedrivs är vid sin publicering anpassad för vetenskapliga tidskrifter. Artiklarna är ofta teoretiska och inomvetenskapligt specialiserade. Det finns emellertid mycket forskning, framför allt empirisk och policyrelevant sådan, som är intressant för en bredare krets. Målet med SNS Analys är att göra denna forskning tillgänglig för beslutsfattare i politik, näringsliv och offentlig förvaltning och bidra till att forskningen når ut i medierna. Finansiellt bidrag har erhållits från Jan Wallanders och Tom Hedelius Stiftelse. Författarna svarar helt och hållet för analys, slutsatser och förslag.

Grafisk form: Patrik Sundström. *Original:* Gabriella Stjärnborg.

Inledning

Historiskt har det funnits en skiljelinje mellan urbanekonomi och städers faktiska fysiska förhållanden. Samhällsvetenskaplig forskning har varit bortkopplad från sådant som arkitektur och gatumiljöer. Det beror bland annat på att det har saknats mätningar av städers fysiska attribut. Den snabbt ökande tillgången på stora datamängder (*big data*) kommer att ändra på detta. Dessa stora datamängder kan potentiellt förvandla ett gaturum till data och möjliggör en större och mer högupplöst bild av det urbana livet än vad som någonsin tidigare funnits. I kombination med prediktiva algoritmer gör dessa stora datamängder att utfallsvariabler såsom bostadspriser och inkomster kan mätas på platser där informationen tidigare varit otillgänglig. Denna rapport visar några exempel på hur stora datamängder kan användas för att utveckla städer.

Potential att brygga gapet mellan urbanekonomi och stadens utseende.

För att besvara klassiska samhällsvetenskapliga frågeställningar som huruvida urban tillväxt och den fysiska staden interagerar med sociala utfall blir stora datamängder ett kraftfullt verktyg när de kombineras med exogena källor till variation. Med exogen menas något som ligger utom en enskild individs kontroll, exempelvis dåligt väder eller en nytt politiskt beslut. I urbana sammanhang är de två huvudkällorna till exogen variation ”platschocker” och ”individchocker”. Platschocker utgör högfrekventa händelser som påverkar geografiska regioner, exempelvis öppnande av stora tillverkningsfabriker.[1] Individchocker är högfrekventa händelser som påverkar geografiska regioner *inom* städer, exempelvis det så kallade ”moving to opportunity”-experimentet.[2] Information om sådana exogena händelser, ”chocker”, gör att observationsstudier får vissa av de fördelar som klassiska statistiska laborationsstudier har, i vilka en forskare tilldelar subjekten en på förhand bestämd ”behandling”. Detta lindrar de problem som annars är vanliga i observationsstudier,[3] och möjliggör för forskaren att särskilja kausalitet från korrelation. Kunskap om kausala effekter är användbart vid framtagande av policyer.

Den bästa förutsättningen ges då stora datamängder används för att studera ekonomiska utfall i städer när högupplösta geografiska data kan matchas med longitudinella data[4] och där exogena händelser kan kopplas till specifika platser. Under sådana förutsättningar gör stora datamängder det möjligt att studera om en policy har en kausal effekt på människor i närheten av den exogena händelsen, oavsett var de flyttar.

Stora datamängder förbättrar också förvaltningen av städer. Genom att göra sina beslut datadrivna i större utsträckning kan städer finjustera regleringar, förbättra allokering av sina knappa resurser och förutspå framtida behov. Det är också vanligt att även själva möjligheten att prediktera framtida utfall är värdefull i sig, utan att det krävs ingående kunskap om de underliggande mekanismerna som orsakar själva utfallen. Vidare är många datadrivna interventioner skalbara; expansionen av datainsamling och digitiseringsinsatser drar till sig entreprenörskap och innovationer.

I en bredare mening är rönen i denna SNS analys relevanta på följande sätt. Politiker och beslutsfattare gynnas av att ha realtidsinformation om hur olika kvarter, grannskap och det lokala samhället utvecklas över tid. Detta är användbart både vid planering och policyskapande. Potentiella investerare och husköpare gynnas också av information om hur exempelvis priser på boende och fastigheter kommer att förändras i olika bostadsområden över tid. Därtill kan fastighetsinvesterare vilja veta var fastighetspriserna sannolikt kommer att stiga. Skalan och magnituden av urbanisering i sig gör dessa frågor relevanta i ett globalt perspektiv.

1. Se exempelvis Greenstone, Hornbeck och Moretti (2010).

2. Chetty, Hendren och Katz (2016); Katz, Kling och Liebman (2001). I detta experiment slumpade USA:s bostads- och stadsplaneringsdepartement under 1990-talet ut kuponger till hushåll i socialt utsatta områden. Kupongerna gav hushållen möjlighet att flyttat till mindre socialt utsatta områden. Chetty, Hendren och Katz kunde senare visa att barn som flyttade till mindre socialt utsatta områden i genomsnitt fick en tredjedel högre inkomst än de som inte flyttade.

3. Ofta kallade endogenitetsproblem, exempelvis omvänd kausalitet, snedvridning orsakad av utelämnade variabler och självselektion.

4. Longitudinella data innebär upprepade mätningar på samma subjekt över tid, exempelvis en individs inkomst under två (eller fler) år.

Tillsammans med mina medförfattare Hyunjin Kim, Scott Duke Kominers, Michael Luca och Nikhil Naik har jag studerat dessa frågor i flera tidigare arbeten.[5] Denna rapport summerar resultaten i dessa studier som baseras på data insamlad i USA. Många rön och resultat kan emellertid antas vara tillämpliga även i en europeisk kontext.

Kvantifiering och stadspolitik

De stora datamängderna och den ökade kvantifieringen av stadsrummet är inte enbart värdefulla för den urbana samhällsvetenskapen. De kan bidra till att dramatiskt förändra sättet städerna fungerar på och avsevärt förbättra möjligheterna att utvärdera den förda politiken.

De större datamängderna påverkar stadspolitiken och förvaltningen på en rad nya sätt: stater digitiserar och delar information och arkiv, företag samlar in högfrekventa mätningar av lokal affärsverksamhet, trafik och andra urbana företeelser. I detta avsnitt presenteras en taxonomi av nya typer av urbana data som nu är tillgängliga för forskare och politiker. Avsnittet avslutas med en diskussion om hur dessa nya data kan påverka den förda politiken.

Taxonomi av datakällor

Digitala utsläpp

En värdefull men underutnyttjad datakälla är *digitala utsläpp*, som består av de digitala avtryck som lämnas online efter en individs vardagsanvändning av internet. Digitala utsläpp kan hjälpa till i mätningar av den fysiska staden inom en mängd områden. Recensionsplattformar, såsom Yelp och TripAdvisor, ger direkta mätningar av kvaliteten på exempelvis service och restauranger i städer världen över. Sociala medier-plattformar, såsom Twitter och Facebook, ger information om pulsen i ett kvarter eller strukturen i sociala nätverk. Linkedin ger information om arbetsmarknad och sökkostnader.[6] Söksträngar från plattformar som Google och Bing ger en inblick i behoven och preferenserna i en fysisk stad. Zillow ger nya insikter om bostadsmarknaden precis som data från delningsplattformar som Airbnb.

Data som genererats av digitala utsläpp kan tillämpas direkt i stadsförvaltningen. Recensioner från Yelp kan exempelvis ge detaljerade högfrekventa data om restauranger som kan användas för att bedöma hygienstandarden[7], och Googlesökningar kan användas för att förutsäga influensautbrott.[8]

Digitala utsläpp kan förbättra staden.

Offentliga handlingar

För trettio år sedan sparade de flesta städer sina uppgifter på papper. Nu håller en bred digitiseringsrörelse på att konvertera data från papper till elektronisk form som kan läsas av maskiner. Denna digitiserade information blir ofta tillgänglig online. Till exempel har brottsregister varit tillgängliga för allmänheten i flera decennier i många amerikanska delstater. Trots detta har det varit svårt för beslutsfattare, forskare och inte minst allmänheten att få tillgång till dem. Över tid har dessa brottsregister digitiserats och gjorts mer tillgängliga, vilket har underlättat för forskning men också förändrat incitamenten för återfallsförbrytare och kriminellt beteende generellt.[9]

En ”öppna data”-rörelse ökar transparensen genom att göra städers interna data tillgängliga för allmänheten. Många större städer (exempelvis Boston, Chicago, Köpenhamn, San Francisco och Stockholm) har skapat öppna ”dataportaler” där forskare och invånare fritt får använda digitiserade uppgifter; många av dessa dataportaler uppdateras i realtid. Tillgängligheten till öppna data uppmuntrar entreprenörer att undersöka vilka nya möjligheter som står till buds och hur en stad kan använda data för att förbättra välfärden samtidigt som den skapar möjligheter för nya samarbeten mellan stadens tjänstemän och forskare.

5. Glaeser m.fl. (2017), Glaeser, Kim och Luca (2017 och 2018).

6. För en överblick av dessa och andra plattformar med användargenererade innehåll, se Luca (2016).

7. Se exemplet nedan och Glaeser m.fl. (2016); Kang m.fl. (2013).

8. Se exempelvis Carneiro och Mylonakis (2009); Ginsberg m.fl. (2009); Polgreen m.fl. (2008); Yang, Santillana och Kou (2015).

9. Finlay (2009); Luca (2016).

Företagsdata
Privata data från företag är ett tredje men mindre utvecklat tillvägagångssätt att mäta den fysiska staden. I tillägg till det som diskuterats ovan kan man föreställa sig att information om medlemskap i träningsanläggningar kan användas för att förstå hälsobeteenden, information i universitetens register för att undersöka studenternas prestationer och data från kreditkortstransaktioner för att kvantifiera spenderandeförändringar över tid. Vidare kan data från telefonoperatörer såsom Vodafone och Telia ge information om i vilka mönster stadens invånare rör sig.

Hur kan nya data stärka staden?

Användarrecensioner kan nyttjas för att allokera stadens resurser.

Egentligen kan man säga att städer är specialiserade på tre aktiviteter som antingen är beroende av eller kan göras betydligt bättre genom data och analys:

1. utvärdera och fastställa policy och regleringar
2. driva offentlig service
3. prognosticera framtida aktivitet i syfte att förbättra planering och arbeta fram ny politik som formar nya riktlinjer.

Härnäst beskriver jag hur man genom att använda stora datamängder kan påverka dessa tre aktiviteter.

Policyutvärdering
Exempelvis går det att studera vilket genomslag en beskattning på hotell har fått: man tar helt enkelt reda på vilket rumspris en hotellbesökare betalat. Det hade varit mer fördelaktigt att studera hur skatterna slår i ett bredare perspektiv, men historiskt sett har andra faktorer som exempelvis effekten av skatter på ”kvalitet” varit väldigt svåra att mäta. Nu kan förändringar i skattesatsen, recensioner i TripAdvisor samt information från Priceline och Airbnb användas för att tillsammans skapa en mycket bredare bild av både de planerade och oplanerade konsekvenserna av skattepolitiken på den fysiska staden.

Förutom att bredda utfallsvariablerna kan nya data möjliggöra en högre frekvens av estimat av förändringar. Anta exempelvis att någon vill utvärdera genomslaget av arbetslöshetsersättning på jobbsökande. Traditionella analyser kanske betraktar hur länge en individ varit arbetslös och hur stor den genomsnittliga inkomsten är efter ett år. Men LinkedIn och liknande sidor skulle i princip kunna ge oss information om individens dagliga jobbsökarbeteende.

Drift av offentliga tjänster
Även om samhällsvetenskaplig forskning traditionellt har fokuserat på effektutvärdering av policyer har en förståelse för den praktiska betydelsen av prediktionsproblem vuxit fram hos forskarna. Städerna är ansvariga för att allokera knappa resurser. Exempelvis väljer staden vilka fall av våld i hemmet som ska följas upp och vilka arbetsrättsliga klaganden som ska utredas. I båda dessa exempel är det underliggande valet inte en utvärdering av policy utan ett prediktionsproblem. Staden måste prediktera vilka förövare som sannolikt kommer att återfalla och vilket arbetsrättsligt fall som kommer att avslöja ett allvarligt arbetsrättsligt problem. Att använda data för att förbättra dessa prediktioner i den fysiska staden skapar värde, och nya datakällor är centrala för denna uppgift.[10]

Ett annat exempel rör teleoperatören Telia som hjälpte lokaltrafiken i Helsingforsregionen att identifiera trängsel i lokaltrafiken för att sätta in matarbussar till tunnelbanan i Helsingfors/Esbotrakten. Detta gjorde pendlingen bekvämare och fler personer valde att resa kollektivt, vilket i sin tur minskade biltrafiken med åtta procent under perioden november 2017 till januari 2018.

Prognostisering
Stadsplanerare och politiker prognosticerar framtida ekonomisk aktivitet genom analyser av tidsseriedata över ledande indikatorer på ekonomisk aktivitet. Nya datakällor, speciellt i kombination med maskininlärning, har potentialen att revolutionera prognostisering. Zillow, TripAdvisor och Linkedin ger mätningar som kan användas i uppskattningar av framtida priser på boende, turism och arbetslöshet. Data

10. Se Kleinberg m.fl. (2015).

från appbaserade betalsystem kan ge insikt om framtida spenderande inom detaljhandel och andra konsumtionsmönster.

Avslutande exempel: Datadrivna hygieninspektioner

Avsnittet avslutas med ett tillämpat och konkret exempel över hur data och prediktionsverktyg kan användas för att direkt informera om hur staden på ett effektivt sätt kan variera allokeringen av sina knappa resurser.

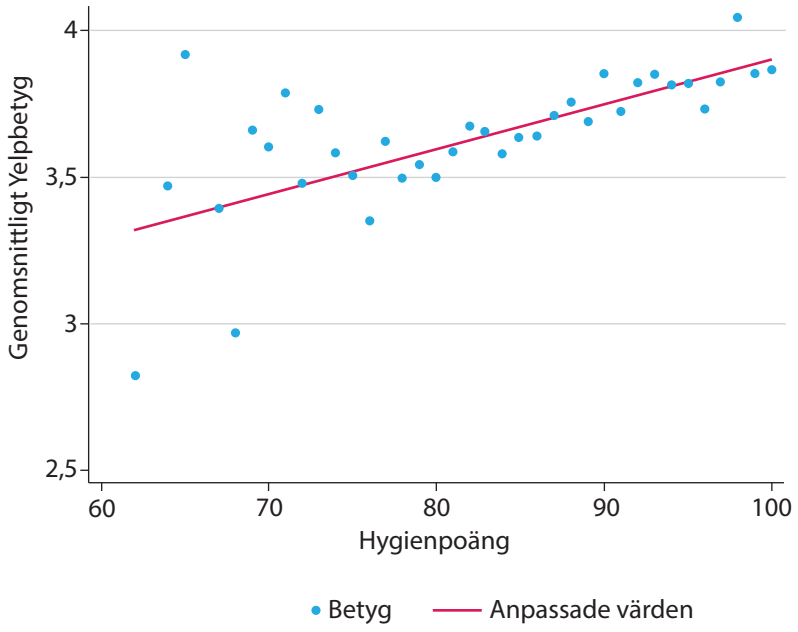
I nästan varje utvecklat land besöker hälsoinspektörer restauranger för att identifiera hälsorisker (exempelvis förvaring av mat i felaktiga temperaturer) som kan leda till matrelaterade sjukdomar. Dessa inspektörer allokeras i regel utifrån hur stor man uppfattar att hälsorisken är hos en restaurang. Kanske är det så att en sushirestaurang inspekteras oftare än en hamburgerrestaurang då inmundigande av sushi har en större sannolikhet att leda till matförgiftning. I övrigt fördelas hälsoinspektörerna i princip slumpmässigt till de olika restaurangerna.

Men inspektionerna *behöver* inte vara slumpmässigt fördelade. Anta i stället att sannolikheten för inspektion skulle baseras på recensioner i Yelp. Kanske genom att söka på orden ”sjuk” eller ”smutsig”. Då skulle man nog hitta en del syndare. Men en prediktiv algoritm tränad med maskininlärning skulle kunna göra mycket mer än så. Algoritmen skulle ”lära sig” från historiken av recensioner – i kombination med information av inspektionsresultat – och sedan prediktera sannolikheten att finna överträdelser baserade på nyare recensioner. Inspektörerna skulle sedan kunna allokeras om till de restauranger som mest troligt överträder hygienbestämmelserna.

Två studier har undersökt möjligheterna att använda språkteknologi (*natural language processing*) för att prediktera brott mot hygienbestämmelser med hjälp av recensioner på Yelp.[11] Även en enkel justering av allokeringen av inspektörer skulle kunna vara mycket kraftfull, vilket illustreras i figur 1. Figuren visar korrelationen mellan betyg givna i Yelp till olika restauranger på ena axeln och restaurangernas hygienpoäng på andra axeln. Även innan tillämpning av maskininlärningstekniker står det klart att Yelpbetyg kan hjälpa till att identifiera hygienpoäng. Staden Boston arrangerade en

11. Kang m.fl. (2013); Glaeser m.fl. (2016).

Figur 1. Korrelation mellan Yelpbetyg och hygieninspektörers poäng.



Not: Recensionsdata bestående av recensioner på Yelp.com av restauranger i San Fransisco från september 2010 till och med september 2013. Hygienpoängen kommer från San Franciscos hälsoskyddskontor för samma period.

tävling för att utveckla prediktiva algoritmer för något av stadens verksamhetsområden. En studie visade att användningen av Yelprecensioner för att styra hälsoinspektioner möjliggör en ordentlig ökning av antalet hittade hälsorisker, utan att kostnaderna för dessa inspektioner ökar.[12]

Men för att svara på frågor av kausal karaktär, exempelvis vilken effekt en specifik policy har, behövs precis som diskuterades i början av den här rapporten också information om exogena chocker. Andra policyer kan dock utvärderas direkt via data, ofta i kombination med prediktiva algoritmer. Detta diskuteras vidare nedan.

Att mäta gatubilden

Bilder kan användas som data ...

I detta avsnitt demonstreras hur datorseendealgoritmer i kombination med bilder från gatuvybilder från Google (*Google Street View)* – eller andra bilder – kan användas (1) för att mäta olika stadsdelars fysiska karaktäristerik och (2) för att estimera ett grannskaps inkomst.

Under det senaste decenniet har Google fotat en stor del av den bebyggda omgivningen i fler än hundra länder. Nästan alla amerikanska städer har dokumenterats i högupplösta bilder och bilderna kan klassificeras med hjälp av datorseende (*computer vision*) algoritmer. En del av dessa bilder kan länkas till gps-kodade attribut av intresse (exempelvis bostadspriser, inkomst eller det allmänna skicket på den urbana miljön). Denna information kan utgöra en träningsdatamängd för att träna algoritmer. Efter träningen visas bilder för algoritmerna som de inte sett förut (detta kallas för utfallsprov, *test sample*) för att avgöra om träningen gick bra. Om så är fallet är det troligt att algoritmen gör goda prediktioner av bostadspriser, inkomster och så vidare utifrån bilder av omgivningar i andra städer. Potentialen i detta tillvägagångssätt demonstreras nedan. Två exempel på proceduren ges i två olika men relaterade frågor.

... för att prediktera inkomster och bostadspriser.

1. Kan gatubilder användas för att prediktera inkomster?
2. Kan gatubilder användas för att förbättra prediktionerna av hur vissa egenskaper i byggnader påverkar bostadspriser?[13]

Den första frågan är mest relevant för utvecklingsländer där det ofta finns stora mängder bilder, men få pålitliga data över inkomster. Den andra frågan är sannolikt relevant i mer utvecklade länder där det finns prisdata att tillgå men där gatubilder inte använts som förklarande variabler. Prediktion av priser med gatubilder har potentiellt ett värde för offentlig politik som ett verktyg för fastighetsvärdering på platser (i både utvecklingsländer och mer utvecklade länder) där fastighetsskatt tas ut.

Om bilder är tillgängliga för en hel stad kan en datorseende modell tränad på ett mindre stickprov av inkomstdata generera en välståndskarta över hela staden likaväl som mått på inkomstsegregationen i staden. Om det finns bilder uppmätta över olika tidpunkter är det möjligt testa hur enskilda insatser ändrar fördelningen av välstånd.

Prediktion av inkomst och bostadspriser med pixlar

Som konceptbevis (*proof of concept*) demonstreras att medianinkomsten bland New Yorks invånare kan predikteras utifrån gatubilder med en datorseende modell. Vidare visas att en sådan modell tränad på bilder tagna i New York klarar av att prediktera medianinkomsten utifrån bilder tagna i Boston med nästan samma precision som för New York. Till sist länkas predikterad inkomst till bostadspriser för att visa potentialen i denna teknik i hedoniska regressioner. Proceduren är mer utförligt beskriven i Glaeser m.fl. (2017), där huvudresultaten kan sammanfattas enligt följande:

12. Glaeser m.fl. (2016).

13. Kan således anpassningen av hedoniska regressioner förbättras? Hedonisk regression är en teknik som används för att avgöra hur olika bostadskaraktäristika påverkar priset på bostaden. Här antas priset på en lägenhet bero på ett antal attribut, till exempel lägenhetens storlek, om det är en vindsvåning eller inte, avståndet till närmaste tunnelbanestation och i vilken del av staden lägenheten finns. Med hedonisk regression kan man uppskatta marknadens värdering av sådana attribut genom att notera hur de i sin tur påverkar lägenhetens marknadspris.

- Den tränade modellen fungerar bra i andra kontexter än tränings- och utfallsprovsdatamängderna.
- Inkomstmåttet som predikterats enbart av bilderna fångar 77 procent av variationen i det sanna inkomstmåttet.
- En modell tränad på bilder tagna i New York fungerar väl för att prediktera inkomster i Boston. Beakta dock att Boston och New York är tämligen lika till utseendet. Vidare är analysen här gjord på kvartersnivå och inte utifrån individuella adresser.

Hedonisk prissättning

Nu vänder vi oss till nästa övning: att länka bilder till priser. I detta fall är vi intresserade av i vilken utsträckning fysiska attribut kan bidra med förmåga att prediktera bostadspriser. I vissa fall kan det finnas ett värde i att enbart öka den prediktiva förmågan hos en modell, exempelvis om regeringen vill förbättra en automatiserad utvärderingsprocess för fastighetsskattrelaterade frågor. I andra fall kanske det är intressant att veta vilka specifika fysiska attribut som förklarar skillnaden i själva bostadspriserna. Gatubilderna kan vara användbara i båda fallen. Exempelvis har datorseende tekniker möjliggjort identifiering av vissa gatunivåattribut såsom väggropar. I princip kan dessa attribut sedan läggas till i en hedonisk regressionsmodell. För tillfället ligger dock fokus på den något enklare uppgiften att prediktera bostadspriser med pixlar.

I huvudsak ämnar vi svara på frågan huruvida bostadsområdens fysiska attribut som attraherar välbeställda personer också ökar själva bostadspriserna. De huvudsakliga resultaten kan summeras enligt följande:

- Den utifrån bilderna predikterade inkomsten förklarar variationen i bostadspriserna utomordentligt, både i träningsdatamängden och i utfallsprovsdatamängden.
- De saker som kan ses från gatan (alltså i bilderna) har ungefär lika stark förmåga att prediktera bostadspriserna som de saker som inte kan ses från gatan.
- Liknande resultat gäller för att prediktera bostadspriserna i Boston med hjälp av en modell tränad på bilder på New York.

Slutsatsen är att gatubilder från Google Street View kan prediktera inkomst i New York (och Boston med modellen som tränats på New York-bilder), och att predikterad inkomst hjälper oss att prediktera bostadspriser i vår datamängd. Detta betyder inte att vi kan prediktera inkomster väl i den utvecklande delen av världen, men det ger hopp om att Google Street View och liknande produkter gör det möjligt att bättre förstå mönster i välstånd och fattigdom världen över.

Nutidsprognosticering av gentrifiering

Gentrifiering kan löst sägas vara en process där nedslitna urbana kvarter renoveras på grund av att relativt välbärgade hushåll flyttar dit. Detta kan leda till ett antal konsekvenser, som att restauranger och närbutiker trängs undan, vilket implicerar att fördelningen av arbetsplatser sorteras om och den etniska sammansättningen ofta förändras. Därför är det ytterst relevant för beslutsfattare och policyer att kunna mäta gentrifieringen. Detta avsnitt föreslår ett sätt att göra det baserat på gatubilder.

Mätningar av förändringar i stadsbilden kan utgöra underlag till beslutsfattare.

Beskrivning av datamaterialet

Vårt första mått på gentrifiering baseras på data från den amerikanska motsvarigheten till Boverket (Federal Housing Finance Agency, FHFA). Dessa data representerar årliga upprepade försäljningsindex från över 18 000 femsiffriga postnummer i USA och beskrivs närmare i Bogin, Doerner, och Larson (kommande). Vi använder information över åren 2012–2016. Den årliga reala tillväxttakten i indexet över denna tidsperiod är 3,1 procentenheter.

Tre mått på demografiska kvarterskaraktäristika används. Dessa är tillgängliga för femårsperioder från en undersökning (American Community Survey, ACS) utförd av den amerikanska motsvarigheten till Statistiska centralbyrån (United States Census Bureau). Måtten är andel med eftergymnasial utbildning (college), andel i åldern 25–34 år och andel vita. Eftersom längden på utbildning tenderar att vara korrelerad med både inkomst och bostadskostnad är

andelen människor med eftergymnasial utbildning i ett område ett rimligt mått på gentrifiering. Vårt stickprov visar att antalet vuxna med en eftergymnasial utbildning ökat med 2,6 procent i ett genomsnittligt postnummerområde i New York.

Vårt sista mått på förändring av grannskapet är förändringen i ”StreetScore”, hämtad från Naik m.fl. (2017). Detta mått innehåller information om hur respondenter bedömt bilder från Google Street View när det gäller upplevd trygghet. Dessa bedömningar har använts som träningsdata för datorseendetekniker, vilka i sin tur genererat StreetScores för ännu fler kvarter. StreetScore tolkas här som en approximation för den allmänna fysiska kvaliteten på ett kvarter snarare än trygghet i sig.

För mått på förändring i företagskategorier använder vi data från Yelps företagsförteckningar som inkommit via användare, företagsägarrapporter, partnerförvärv samt interna kvalitetskontroller.[14]

Resultat på lokala bostadspriser

Först undersöker vi förmågan hos Yelpdata att prediktera samtida förändringar i prisökningar på bostäder på postnummernivå under perioden 2012–2016 på ett liknande sätt som Rascoff och Humphries (2015), vilka länkar ett postnummers närhet till en Starbucks och prisökningen i Zillow. I vår version undersöker vi om prisökningen är korrelerad med en ökning av antalet Starbuckscaféer som då visar om ytterligare ett Starbuckscafé är en indikator på gentrifiering.

Först beräknas korrelationen mellan ökningen i bostadspriser och ökningen i antalet Starbuckscaféer i samma postnummerområde och samma år. Ytterligare ett Starbuckscafé ett givet år är associerat till en ökning i bostadspriser på ungefär 0,5 procent. Denna effekt är stor, men förklaringsgraden hos modellen är inte speciellt bra. Vidare är det oklart åt vilket håll kausaliteten går. Det skulle ju kunna vara så att Starbucks tenderar att förlägga nya caféer till områden som är på uppgång. I detta fall skulle då korrelationen spegla Starbucks affärsstrategi, alltså att en hög tillväxt i bostadspriser i ett givet postnummerområde under ett givet år orsakar att Starbucks öppnar café där. Å andra sidan kan företag bidra till gentrifiering. I ett sådant scenario skulle en nyöppnad Starbucks göra att människor med högre genomsnittlig inkomst skulle flytta till postnummerområdet där detta café öppnade. Här är kausaliteten alltså omvänd i jämförelse med det förra exemplet.

För att delvis skilja på dessa potentiella scenarier kontrollerar vi för icke-observerade olikheter mellan postnummerområden som inte varierar över tid. Tidsperioden vi betraktar är kort, så om Starbucks tar sikte på tillväxtområden skulle detta förfaringssätt ta bort mycket av korrelationen. I detta fall minskar den 0,5 procentiga ökningen till en 0,17 procentig ökning av bostadspriserna i postnummerområden där antalet Starbuckscaféer ökar med en enhet.

En annan analys inkluderar både mått på nuvarande såväl som tidigare tillväxt i antalet Starbucks, och antalet Starbucks som stängs och därtill antalet recensioner av Starbucks. Ökningen i antalet Starbucksrecensioner har en prediktiv förmåga på förändringar i kvarteret; en ökning i antalet recensioner av Starbucks är associerad med en 1,4 procentig ökning i bostadspriserna inom ett givet postnummer. Eftersom förekomsten av ett Starbuckscafé är mindre viktig i jämförelse med recensionerna av Starbucks (1,4 är större än både 0,5 och 0,17) ifrågasätts förklaringen att människor betalar för att flytta nära en Starbucks. Samtidigt som Starbucks kanske är en framträdande cafékedja är det inte den enda aktören inom detaljhandeln som signalerar gentrifiering på lokal nivå. Analysen utökas därför till att inkludera alla möjliga caféer listade hos Yelp under samma tidsperiod. Detta utökar antalet postnummerområden eftersom flera postnummerområden har åtminstone ett café under tidsperioden vi studerar. Liknande resultat hittas, även om

14. Trots dess höga upplösning och goda tillgänglighet har Yelpdata begränsningar, vilka diskuteras mer ingående i Glaeser, Kim och Luca (2017). Yelps företagsklassifikationer tilldelas via användar- och företagsägarrapporter, vilka ofta resulterar i en icke-systematisk näringsgrenskategorisering som inte korresponderar mot offentligt producerade datamängder. Vidare beror kvaliteten på Yelpdata på graden av Yelpanvändande, vilket har ökat över tid. Här klassificerar vi företag som aktiva om de har mottagit minst en Yelprecension.

magnituden i ökningen på bostadspriser som är associerad till recensioner av caféer minskar något. Skillnaden mellan Starbucks och det utvidgade cafémåttet ger ett visst stöd till hypotesen att människor med högre genomsnittlig inkomst flyttar till postnummerområden där caféer öppnar.

I Glaeser, Kim och Luca (2017) utökar vi analysen till att innehålla andra näringsgrenar som finns klassificerade i Yelp. I många fall som liknar Starbucksexemplet ökar antalet recensioner i Yelp den prediktiva förmågan utöver den som innehåller enbart en ökning av antalet företag, vilket tyder på att både förändringar i den lokala ekonomin och i användningen av Yelp är relaterade till gentrifiering.

Resultat kring demografiska förändringar

Här undersöks om det lokala företagsekosystemet förändras med en demografisk förändring i ett grannskap. Fokus ligger på New York och vi studerar demografiska förändringar mellan tidsperioderna 2007–2011 och 2012–2016 och därtill förändringar i StreetScore under perioden 2007–2014.

Förändringen i antalet livsmedelsaffärer är statistiskt signifikant korrelerad med förändringar i antalet vuxna med eftergymnasial utbildning. Korrelationen mellan förändringen i antalet livsmedelsaffärer med ålder och etnisk sammansättning (*racial composition*) inom ett postnummerområde är också statistiskt signifikant, men denna korrelation är ungefär hälften så stor som korrelationen med antalet vuxna med eftergymnasial utbildning. Dessa resultat ligger i linje med forskningen om ”matöknar” (*food deserts*) som syftar på att många mindre ekonomiskt bemedlade människor bor i områden med ett dåligt utbud av hälsosamma matalternativ.

Antalet tvättomater är korrelerat med andelen invånare som är unga, vilket möjligtvis inte är överraskande. Eftersom tvättomater sällan anses ”exklusiva” verkar detta resultat mer kompatibelt med affärsförtätning.

Det finns korrelationer mellan förändringen i andelen av populationen med eftergymnasial utbildning och förändringar i antalet caféer, barer, restauranger, barberare, vinbarer, närbutiker, snabbmatsrestauranger, florister och restauranger vilka av Yelp är kategoriserade som dyra. Restauranger, barberare och florister korrelerar också med antalet människor som är unga. Korrelationen med etnisk sammansättning är svagare i nästan alla av fallen.

I Glaeser, Kim och Luca (2017) reproducerar vi dessa resultat för Boston, Chicago, Los Angeles och San Francisco samt undersöker korrelationen med antalet Yelprecensioner i olika kategorier. Många av mönstren är i stort sett lika, med två större skillnader. Antalet tvättomater är inte längre starkt korrelerat med gentrifiering. I de övriga fyra städerna, till skillnad från New York, korrelerar många av Yelprecensionsantal starkt med antalet yngre människor inom ett postnummerområde; en potentiell förklaring är den geografiska variationen i åldern på Yelprecensenterna.

Det sista måttet är fysisk förändring i ett grannskap mätt genom StreetScore. Som tidigare börjar vi med New York och betraktar sedan andra städer.[15] För att kunna jämföra resultaten fortsätter vi att använda data på postnummernivå, även om det inte finns någon egentlig anledning att inte titta på information på kvartersnivå. På postnummernivå är den starkaste korrelationen med antalet vegetariska restauranger. Korrelationen var mycket svagare mellan vegetariska restauranger och förändringen i andelen eftergymnasialt utbildade. Den näst starkaste korrelationen är med förändringar i antalet Starbuckscaféer, och den tredje starkaste är den med vinbarer. Detta speglar tidigare resultat med avseende på demografiska förändringar.

Avslutande diskussion

Stora datamängder är särskilt värdefulla när de direkt kan förbättra beslutsfattande. Sätten som Yelp (och andra liknande betygsplattformar med användargenererade omdömen) kan förbättra en stads tjänster som exempelvis sanitära inspektioner. Generellt ger stora datamängder samhället bättre beslutsunderlag då kostnaderna sänks för medborgarna att bidra till stadens tjänster. Appar i telefoner och

Stora datamängder kan förbättra beslutsfattandet.

15. Glaeser, Kim och Luca (2017).

liknande förser medborgarna med verktyg de kan använda för att ge återkoppling till staden snabbt och billigt. Denna rapport har visat på potentialen i att mäta gentrifiering, bostadspriser och hur medborgarnas inkomster kan mätas.

Generellt kan stora datamängder förbättra både forskningen kring dem, men bara om de används på ett eftertänksamt sätt. Stora datamängder har mycket större potential för stadsforskning om de kopplas ihop med exogena källor till variation och mycket större potential för beslutsfattande och implementering om de kopplas ihop med en öppenhet gentemot nya metoder.

Diskussioner om hur folkräkningen kan uppdateras eller ersättas har pågått under de senaste åren. Exempelvis har Storbritannien övervägt att ersätta folkräkningen med administrativa data och data från tredje part som exempelvis sökmotorer såsom Google.[16] Vart tionde år genomför USA:s folkbokföringsmyndighet en folkräkning för att bland annat bestämma mandatfördelningen över delstaterna i representanthuset. Denna folkräkning kostar ungefär 20 miljarder dollar, men nu räknar man med att kunna spara 5,2 miljarder av dessa genom att använda administrativa data i kombination med data från tredje part.[17]

Tredjepartsdata kan förbättra officiell statistik.

Analyser av den potentiella datakällan Yelp indikerar att dessa nya datakällor kan vara användbara komplement till officiellt producerad statistik. Information från Yelp kan hjälpa till med att prediktera samtida förändringar i den lokala ekonomin. Yelp kan också ge en direkt bild av en potentiell ekonomisk förändring på lokal nivå. Denna datakälla är ett användbart tillskott till de dataverktyg som lokala beslutsfattare har tillgång till.

Data från plattformar såsom Yelp, i kombination med officiellt producerad statistik, kan ge värdefulla kompletterande dataset med mer lägliga och bättre upplösta prognoser och policyanalyser med en större mängd variabler och på det stora hela en mer komplett bild av den lokala ekonomin.

16. Hope (2010); Sanghani (2013).
17. U.S. Census Bureau (2015); Mervis (2017).

En viktig fråga i relation till allt detta är hur avvägningen mellan individers integritet och öppenheten av den rikliga informationen som finns i datakällor som diskuterats här ska hanteras. Detta är särskilt relevant i ljuset av dataskyddsförordningen (DSF), mer känd som GDPR (efter General Data Protection Regulation) som implementerades i EU under 2018.

Därtill är en viktig fråga den om hur och i vilken utsträckning officiella statistikproducenter såsom SCB ska hantera och/eller komplettera med data genererade av privata aktörer.

Referenser

Bogin, A. N., W. M. Doerner och W. D. Larson. Under utgivning. ”Local House Price Paths: Accelerations, Declines, and Recoveries”, *Journal of Real Estate Finance and Economics*

Carneiro, H. A. och E. Mylonakis (2009). ”Google Trends: A Web- Based Tool for Real-Time Surveillance of Disease Outbreaks”, *Clinical Infectious Diseases*, 49(10): 1557–1564.

Chetty, R., N. Hendren och L. F. Katz (2016). ”The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment”, *American Economic Review*, 106(4): 855–902.

Finlay, K. (2009). ”Effect of Employer Access to Criminal History Data on the Labor Market Outcomes of Ex-Offenders and Non-Offenders” i *Studies of Labor Market Intermediation*, red. D. H. Autor. Chicago: University of Chicago Press, 89–125.

Ginsberg, J., M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski och L. Brilliant (2009). ”Detecting Influenza Epidemics Using Search Engine Query Data”, *Nature*, 457(7232): 1012–1014.

Glaeser, E. L., S. D. Kominers, M. Luca och N. Naik (2017). ”Big Data and Big Cities: The Promises and Limitations of Improved Measures of Urban Life”, *Economic Inquiry*, 56(1), 118–137.

Glaeser, E. L., A. Hillis, S. D. Kominers och

M. Luca (2016). ”Crowdsourcing City Government: Using Tournaments to Improve Inspection Accuracy”, *American Economic Review,* 106(5), 114–118.

Glaeser, E. L., Hyunjin Kim och M. Luca. (2017). ”Nowcasting the Local Economy: Using Yelp Data to Measure Economic Activity”, National Bureau of Economic Research Working Paper 24010.

Glaeser, E. L., Hyunjin Kim och Michael Luca (2018). ”Nowcasting Gentrification: Using Yelp Data to Quantify Neighborhood Change”, *AEA Papers and Proceedings*, 108: 77–82.

Greenstone, M., R. Hornbeck och E. Moretti (2010). ”Identifying Agglomeration Spillovers: Evidence from Winners and Losers of Large Plant Openings”, *Journal of Political Economy*, 118(3): 536–598.

Hope, C. (2010). ”National Census to be axed after 200 years”, The Telegraph, July 9, 2010. <http://www.telegraph.co.uk/news/politics/7882774/National-census-to-be-axed-after-200-years.html>. Besökt 6 juli 2017.

Kang, J. S., P. Kuznetsova, M. Luca och Y. Choi (2013). ”Where *Not* to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews” i *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA: ACL, 1443–1448.

Katz, L. F., J. R. Kling och J. B. Liebman (2001). ”Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment.” *Quarterly Journal of Economics*, 116: 607–651.

Kleinberg, J., J. Ludwig, S. Mullainathan och Z. Obermeyer (2015). ”Prediction Policy Problems”, *American Economic Review*, 105(5): 491–495.

Luca, M. (2016). ”User-Generated Content and Social Media” i *Handbook of Media Economics*, red. S. Anderson, J. Waldfogel, och D. Strömberg. Amsterdam, The Netherlands: North Holland.

Mervis, J. (2017). ”Scientists fear pending attack on federal statistics collection”, *Science Magazine*, 3 januari 2017. <http://www.sciencemag.org/news/2017/01/scientists-fear-pending-attack-federal-statistics-collection>. Besökt 6 juli, 2017.

Naik, N., S. Duke Kominers, R. Raskar, E. L. Glaeser och C. A. Hidalgo (2017). ”Computer Vision Uncovers Predictors of Physical Urban Change”, *Proceedings of the National Academy of Sciences* 114 (29): 7571–7576.

Polgreen, P. M., Y. Chen, D. M. Pennock, F. D. Nelson och R. A. Weinstein (2008). ”Using Internet Searches for Influenza Surveillance”, *Clinical Infectious Diseases*, 47(11): 1443–1448.

Rascoff, S. och S. Humphries (2015). ”Confirmed: Starbucks Knows the Next Hot Neighborhood Before Everybody Else Does”, *Quartz*, January 28, 2015. https://qz.com/334269/what-starbucks-has-done-toamerican-home-values/. Besökt 4 januari 2018.

Sanghani, R. (2013). ”Google could replace national census”, *The Telegraph*, 26 juni 2013. <http://www.telegraph.co.uk/technology/google/10142641/Google-could-replace-national-census.html>. Besökt 6 juli 2017.

U.S. Census Bureau (2015). ”2020 Census Operational Plan Overview and Operational Areas”, <https://censusproject.files.wordpress.com/2015/12/2020-census-opplan-conference-call_the-census-project_10-21-15_final-1.pdf>. Besökt 6 juli 2017.

Yang, S., M. Santillana och S. C. Kou (2015). ”Accurate Estimation of Influenza Epidemics Using Google Search Data via ARGO”, *Proceedings of the National Academy of Sciences of the United States of America*, 112(47): 14473–14478.