UPTEC W 20013
Examensarbete 30 hp
Mars 2020

UPPSALA
UNIVERSITET

En utveckling av alternativa metoder för klassificering av avrinningsområden med avseende på bebyggelsetyp och anslutningsförhållanden

William Bredberg

REFERAT

En utveckling av alternativa metoder för klassificering av avrinningsområden med avseende på bebyggelsetyp och anslutningsförhållanden
William Bredberg

Olika hårdlagda ytor och olika kombinationer av ledningssystem ger upphov till olika avrinningskoefficienter. Klassificering av avrinningsområden med avseende på bebyggelsetyp och anslutningsförhållanden används därför ofta i syfte att förenkla avrinningsberäkningar. Arbetet det innebär att utföra dessa klassificeringar är ofta tidsödande när det utförs manuellt. Därför ämnade examensarbetet att förenkla detta arbete.

I examensarbetet utvecklades tre alternativa metoder för klassificering med avseende på bebyggelsetyp, samt en metod för klassificering av anslutningsförhållanden. De två första metoderna ämnade för bebyggelstyper omfattade rektangulärklassificering och klassificering baserad på största sannolikhet. Med dessa var ett parameterval och träningsdata nödvändigt. Den tredje metoden baserades på närmaste granne-algoritmen och krävde endast träningsdata i form av att ett visst antal områden hade klassificerats på förhand. Dessa tre första metoder utvärderades mot tidigare utförda klassificeringar med hjälp av förväxlingsmatriser. Metoden för klassificering av anslutningsförhållanden utvecklades utifrån ett givet antal villkor. Den metoden hade inte samma utvärderingsmöjligheter och validerades därför genom inspektion av exempelområden.

För bebyggelsetypsklassificering uppvisade metoden baserad på närmaste granne-algoritmen lovande resultat med en överensstämmelsegrad som var nästintill perfekt under vissa förutsättningar. Metoden begränsas dock av städers utformning och områdesfördelning, varför användandet av metoden kräver en förståelse för hur den fungerar och under vilka förutsättningar den kan tänkas prestera sämre. Samtliga andra metoder ämnade för bebyggelsetypsklassificering gav otillräckliga resultat med de parameterval som testades. Metoden för klassificering av anslutningsförhållanden följde givna villkor och föreföll prestera väl vid inspektion av enskilda exempelområden. Dataunderlaget kan däremot utgöra en begränsning och metoden hade behövts utvärderas i flera städer för att skapa ett bättre underlag för utvärdering.

Nyckelord: anslutningsförhållanden, avloppsnät, avrinningsområde, bebyggelsetyp, dagvatten, FME, hydraulisk modellering, klassificering, maskininlärning, spillvatten

Institutionen för geovetenskaper, Luft- vatten- och landskapslära, Uppsala universitet
Villavägen 16, SE-752 36 Uppsala

ABSTRACT

A Development of Alternative Methods for Catchment Area Classification with Regards to Building Type and Water Pipeline Connection
William Bredberg

Surfaces of varying permeability and different combinations of water pipelines within catchment areas result in different runoff coefficients. Classification of catchment areas with regards to building type and water pipeline connections are therefore commonly used in order to simplify runoff calculations. The work required to perform these classifications is often time consuming as it is performed manually. This master's thesis aims to make the task of classification less labour intensive.

Three alternative methods for classification with regards to building type and one method for water pipeline connection were considered in this master's thesis. The first two building type classification methods were based on parallellepiped and maximum likelihood classification theory and required a choice of parameters and training data in order to perform classifications. The third method was based on a nearest neighbour algorithm and only required training data in the form of a number of previously classified areas. All three building type classification methods were evaluated against previously obtained classifications for catchment areas using confusion matrices. The method purposed for catchment area water pipeline connections was developed from a given number of predetermined conditions. That method did not have the same evaluation opportunities as no previously made classifications could be obtained. Instead, the method was validated by inspection of individual example areas.

The method based on the nearest neighbour algorithm proved to perform very well, with a strength of agreement that was almost perfect under certain conditions. The performance of the method was however limited by city formation and area distribution. Hence, future usage of the method require an understanding of how the algorithm performs the classifications and under what circumstances the method may or may not produce sufficient results. All other methods purposed for building type classification produced inferior results with the parameters that were tested. The water pipeline connection classification method followed the outlined conditions and seemed to perform well. It is however subject to data limitations and the method would have to be validated in additional cities on several data sets in order to provide a better basis for evaluation.

Keywords: building type, catchment, classification, sewage system, FME, hydraulic modelling, machine learning, stormwater, wastewater, water pipeline connections.

Department of Earth Sciences, Program for Air, Water and Landscape Science, Uppsala University, Villavägen 16, SE-752 36 Uppsala
ISSN 1401-5765

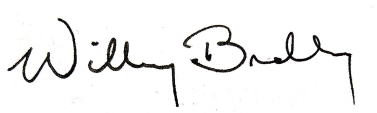
FÖRORD

Detta examensarbetet omfattar 30 högskolepoäng och utfördes som ett slutgiltigt moment inom civilingenjörsprogrammet i miljö- och vattenteknik vid Uppsala universitet och Sveriges lantbruksuniversitet. Uppdraget har utförts på initiativ av Tyréns AB i Stockholm, för deras avdelning för dagvatten och modellering med Hans Hammarlund som handledare för projektet. Rickard Pettersson vid institutionen för geovetenskaper vid Uppsala universitet har varit ämnesgranskare och Björn Claremar vid institutionen för geovetenskaper vid Uppsala universitet har varit examinator.

Ett stort tack riktas till Rickard Petterson för många av de idéer som har legat till stor grund för att ge arbetet riktning. Ytterligare ett stort tack riktas till Hans Hammarlund för hans tålamod, insikt och ingående förklaringar av skillnader i ledningssystem och dess betydelse. Även Martin Rosén på Tyréns AB ska ha ett stort tack för att ha gett mig tips och vägledning i FME desktop som har varit till stor hjälp och har sparat mig mycket tid. Jag vill även tacka Linköping kommun och Västervik kommun som har gett mig tillgång till underlaget som gjorde utvärderingen av mina metoder möjlig.

Slutligen, ett varmt tack till mina nära och kära som har stöttat mig genom mina studier.

Uppsala, december 2019



William Bredberg

UPTEC W 20 013, ISSN 1401-5765.
Digitalt publicerad vid Institutionen för geovetenskaper,
Uppsala universitet, Uppsala, 2020.

POPULÄRVETENSKAPLIG SAMMANFATTNING

När samhället förändras uppstår även behovet för förändringar i våra arbetsmetoder när det kommer till att forma vår omgivning. I takt med att städer växer och nya områden utformas så är det många aspekter som måste tas hänsyn till - bland annat vädrets makter. Det är här hydrauliska modeller kommer in i bilden. Dessa används för att till exempel simulera flödesbelastningar på ledningsnät genom att uppskatta hur mycket avrinning som kan uppstå inom ett område. Avrinning är det vatten som inte infiltreras i marken utan rör sig vidare i ett område efter nederbörd eller när is och snö smälter. På så sätt är det möjligt att skapa beslutsunderlag och använda resurser där de behövs som mest, samt dimensionera strukturer och ledningsnät på bästa möjliga sätt. Därmed blir hydrauliska modeller användbara för till exempel skyfalls- och klimatanpassningsåtgärder.

Tyvärr finns det ett problem med alla typer av modeller. Det är i princip omöjligt att med hjälp av en modell representera verkligheten med alla dess oförutsägbara, komplexa och dynamiska system. Detta mynnar ut i en fråga alla som arbetar med modeller måste ställa sig om och om igen. Hur komplicerad måste en modell vara för att prestera tillräckligt bra? Ju mer information som finns tillgänglig, ju mer detaljerad och avancerad kan en modell bli. Det innebär däremot även att mer arbete krävs för att skapa underlag och mer datorkraft behövs för att köra modellen.

När nya områden uppstår och städer expanderar så finns oftast inte all information tillgänglig på förhand när avloppssystem och ledningsnät ska utformas. Den hydrauliska modelleringen blir därmed en utmaning. Därför är det vanligt att utgå från erfarenhetsvärden för hur mycket avrinning som kan uppstå inom en särskild typ av område. Av den anledningen är det vanligt att områden klassificeras utefter vilken typ av bebyggelse och ledningsnät som finns inom dem, eller kommer att finnas inom dem. Detta fungerar eftersom det redan finns kunskap om hur mycket avrinning som kan uppstå inom ett industri- eller villaområde av en viss storlek. Avrinningen kommer även att påverkas beroende på om det ligger äldre eller nyare ledningsnät inom området. Historiskt sett i Sverige så separerades inte ledningar för smutsigt spillvatten från toaletter och tvätt som idag går till reningsverket och dagvatten från till exempel nederbörd som inte har samma reningsbehov. Ledningsnäten i många områden byggs om och idag så separeras spill- och dagvatten av båda ekonomiska och hygieniska skäl. Därför är det även av intresse att ha lättillgänglig information över vilka kombinationer av ledningssystem och anslutningsförhållanden som finns inom olika områden. Här uppstår nästa problem. För att beräkna avrinning inom stora städer eller stadsdelar blir det ofta nödvändigt att klassificera flera tusen områden med avseende på bebyggelsetyp och anslutningsförhållanden. Detta är något som vanligtvis görs manuellt genom inspektion av enskilda områden och tar därför väldigt lång tid. Detta examensarbetet ämnar att förenkla den tidskrävande och tråkiga arbetsprocessen genom att endast använda vanligt förekommande kartmaterial.

Flera metoder utvecklades och jämfördes, men framförallt två metoder stack ut med lovande resultat. Den ena bygger på att klassificera områden beroende på vad för bebyggelsetyp intilliggande områden har. En algoritm undersöker vilka bebyggelsetyper som finns i närheten av varje område som ska klassificeras och utför sedan klassificeringen

baserat på vilken bebyggelsetyp som är vanligast inom ett optimerat avstånd. Det optimerade avståndet är nödvändigt för att algoritmen inte ska använda områden som ligger för långt bort. Istället för att manuellt klassificera flera tusen områden skulle det alltså endast vara nödvändigt att klassa ett fåtal för att algoritmen skulle få tillräcklig information att arbeta med. Denna metoden uppvisade en nästintill perfekt överensstämmelsegrad med verkligheten eftersom den kunde uppnå mer än 90 % korrekta klassificeringar under vissa förhållanden. Anledningen till att detta fungerar är att städer ofta är planerade och utformade med olika bebyggelsetyper i kluster. Det är trots allt ovanligt att hitta små villor mitt i större industriområden. Därför kan klassificeringarna utföras baserat på vad som finns i närheten av varje område. Givetvis innebär metoden en begränsning beroende på hur en stad eller stadsdel är utformad, varför det är nödvändigt att ha kunskap om när metoden skulle kunna tänkas producera ett sämre resultat. Den andra fungerande metoden klassificerar områden baserat på vilka typer av ledningar som finns inom ett område och ett antal villkor. På så sätt går det att automatiskt erhålla kunskap om vilka områden som har kombinerade eller separerade ledningssystem som påverkar avrinningen. Detta tycks fungera mycket bra vid översiktlig inspektion. Metoden kan tyvärr begränsas av kartmaterialet som finns tillgängligt. Ibland är inte alla ledningar ritade på ett helt korrekt sätt som representerar verkligheten. Algoritmen i metoden processar och manipulerar kartmaterialet genom att till exempel förlänga vissa ledningar för ett bättre resultat, men det finns ingen garanti för att alla områden alltid blir korrekt klassificerade.

I ett försök att hitta skillnader mellan de olika bebyggelsetyperna utvecklades även metoder som separerar de olika klasserna åt baserat på ett antal egenskaper i form av parametrar. Dessa parametrar utgjordes av exempelvis storleken på byggnader inom området eller antalet byggnader i förhållande till områdets area. Dessa metoder uppvisade tyvärr inte lika lovande resultat med de parametrar som testades. Däremot kvarstår möjligheten att finna parametrar som skiljer på olika bebyggelsetyper eftersom att metoden nu redan är utformad. Detta skulle innebära att områden inte behöver klassificeras på förhand, utan att endast kartmaterialet vore nödvändigt.

Modernisering och teknologiska framsteg innebär stora möjligheter när det kommer till att påverka omgivningen. Metoder och sanningar som upptäcktes för många år sedan används fortfarande, men blir mer och mer sofistikerade. Detsamma gäller för modellering och klassificeringsmetoder. Examensarbetet visar på att det finns goda förutsättningar för att använda alternativa metoder för områdesklassificering, men har fortfarande bara skrapat på ytan. Det finns alltid stort utrymme för vidare utveckling och förbättring.

ORDLISTA

Anslutningsförhållande: Servis- och ledningsförekomst och anslutningar inom ett område och intilliggande gata.

Avloppsvatten: Omfattar spill-, drän-, och dagvatten.

Avrinning: Det regn- och smältvatten som rör sig vidare inom ett område.

Avrinningskoefficient: Andelen nederbörd som blir till avrinning.

Dagvatten: Vatten härstammande från nederbörd eller till exempel framträngande grundvatten som avrinner från tak och andra ytor.

Dagvattenservis: Anslutning för dagvatten till ett ledningsnätverk.

Delvis kombinerat system Ledningssystem i vilket dag- och spillvatten från ett område eller en fastighet avleds i samma ledning, men dagvatten från intilliggande gata avleds i en separat ledning.

Dränvatten: Vatten som har dränerats från till exempel husgrunder eller genom annan typ av markavvattning.

Kombinerat system Ledningssystem i vilket dag- och spillvatten från ett område eller en fastighet, samt dagvatten från intilliggande gata avleds i en och samma ledning.

Normalfördelad: En variabel som oftast antar värden nära medelvärdet av samtliga värden utan större avvikelser.

Område: Avser ett avrinningsområde hur det än är definierat. Kan i rapporten ibland avse fastighetsområden.

Separerat system Ledningssystem i vilket spill- och dagvatten alltid avleds i separata ledningar.

Spillvatten: vatten från fastigheter som leds till avloppsledningsnät. Detta kan bestå av förorenat vatten från toaletter, bad, disk och tvätt, samt processvatten från industrier.

Spillvattenservis: Anslutning för spillvatten till ett ledningsnätverk.

Överensstämmelsegrad: Hur väl en en metods resultat stämmer mot verkligheten.

Innehållsförteckning

Referat I
Abstract II
Förord III
Populärvetenskaplig sammanfattning IV
Ordlista VI

1 Inledning **1**
1.1 Syfte och mål 2
1.2 Avgränsningar 2

2 Teori **3**
2.1 Hydraulisk modellering och avloppssystem 3
2.1.1 Flöden i ett avrinningsområde 4
2.1.2 Vanliga beräkningsmetoder för bidragande andel yta 5
2.1.3 Områdesklassificering med avseende på bebyggelsetyp och anslutningsförhållanden 7
2.2 Bildklassificering 10
2.2.1 Rektangulärklassificering 10
2.2.2 Klassificering baserad på största sannolikhet 11
2.3 Klassificering baserad på intilliggande områden 13
2.3.1 Närmaste granne-algoritmen 13
2.4 Statistiska jämförelser 14
2.4.1 Utvärdering med förväxlingsmatriser 14
2.4.2 Cohens kappa 16
2.4.3 Shapiro-Wilks test 17
2.5 Modellportabilitet 18

3 Material och metod **19**
3.1 Områdesbeskrivningar och dataunderlag 19
3.1.1 Linköping 19
3.1.2 Västervik 21
3.2 Analysverktyg 23
3.2.1 ArcGIS 23
3.2.2 R och RStudio 23
3.2.3 FME Desktop 23
3.3 Metod I - Rektangulärklassificering och byggnadsytor 24

3.3.1 Klassificeringsvillkor . . . 24
3.3.2 Klassificeringsmetod . . . 25
3.4 Metod II - Klassificering baserad på största sannolikhet . . . 27
3.4.1 Förberedelse av dataunderlag och parametrar baserade på byggnadsytor . . . 29
3.4.2 Klassificering . . . 29
3.4.3 Konversion . . . 30
3.4.4 Parametrar baserade på servisförekomst . . . 30
3.5 Metod III - Närmaste granne-algoritmen och intilliggande områden . . . 31
3.5.1 Algoritm utan distansviktning . . . 31
3.5.2 Algoritm med distansviktning . . . 33
3.6 Metod IV - Klassificering av anslutningsförhållanden . . . 36
3.7 Utvärdering och jämförelse av metoder . . . 39
3.7.1 Export av prestationsdata . . . 39
3.7.2 Utvärderingsmall för metoder i form av förväxlingsmatris . . . 41
3.7.3 Portabilitet . . . 42
3.7.4 Tidsåtgång . . . 42

4 Resultat **43**
4.1 Metod I - Rektangulärklassificering och byggnadsytor . . . 43
4.2 Metod II - Klassificering baserad på största sannolikhet . . . 44
4.2.1 Parameterval baserade på byggnadsytor . . . 44
4.2.2 Parameterval baserade på servisförekomst . . . 46
4.3 Metod III - Närmaste granne-algoritmen och intilliggande områden . . . 49
4.3.1 Algoritm utan distansviktning . . . 49
4.3.2 Algoritm med distansviktning . . . 52
4.4 Metod IV - Klassificering av anslutningsförhållanden . . . 53
4.5 Portabilitet . . . 58
4.5.1 Närmaste granne-algoritmen . . . 58

5 Diskussion **61**
5.1 Metod I - Rektangulärklassificering och byggnadsytor . . . 61
5.2 Metod II - Klassificering baserad på största sannolikhet . . . 61
5.3 Metod III - Närmaste granne-algoritmen och intilliggande områden . . . 62
5.4 Metod IV - Klassificering av anslutningsförhållanden . . . 63
5.5 Vidare studier . . . 65

6 Slutsatser **67**

7 Referenser **68**

Bilagor **71**

Bilaga A - Transformatorer i FME Desktop 71
Bilaga B - Rektangulärklassificering baserad på byggnadsytor 72
Bilaga C - Klassificering baserad på största sannolikhet 73
Bilaga D - Närmaste granne-algoritmen och intilliggande områden 74
Bilaga E - Export av prestationsdata 77

1 INLEDNING

Hydraulisk modellering har många användningsområden och används som beslutsunderlag i ett flertal processer samt för att modellera ytavrinning. Till exempel kan den användas i syfte att simulera beslastningen på ett dagvattenledningsnät, eller för att utvärdera skyfalls- och klimatanpassningsåtgärder (Blomquist m. fl., 2016). Modelleringen möjliggör även utvärderandet av åtgärdseffekter på ledningsnät och utbyggnad av nya områden redan innan de implementeras. Vidare, så kan dessa åtgärdseffekter även undersökas, optimeras, samt enkelt göras tillgängliga för relevanta beslutsfattare. Därför finns det ett stort behov av att utreda och samla in så mycket information om områden av intresse som möjligt. Vid hydraulisk modellering är framförallt vattenflöden och avrinning av intresse. Avrinningen är det vatten från nederbörd i form av regn och snö som inte avdunstar, utan rinner vidare i landskapet (SMHI, 2017). Det kan även bestå av framträngande grundvatten. Eftersom att alla avrinningsområden är utsatta för nederbördspåverkan finns det ett behov av att uppskatta avrinningen inom dessa områden, då avrinningen är avgörande för hur stora flöden som uppstår inom området. Detta är i sin tur avgörande för till exempel dimensionering av ledningsnät och annan områdesutveckling.

Dock förutsätter modelleringen god områdeskännedom där den implementeras och begränsas bland annat av dess dataunderlag. För att beräkna dessa flöden används ofta en avrinningskoefficient. Avrinningskoefficienten avgör hur stor del av nederbörden som avrinner inom området efter förluster i form av evapotranspiration och magasinering i markytans eventuella ojämnheter (Svenskt Vatten, 2010). Storleken på den del av nederbörden som blir till avrinning beror främst på exploateringsgraden och andelen hårdgjord yta inom området. Vidare, beror det även på områdets lutning, samt regnintensiteten. Om ett område har omfattande bebyggelse kommer därför den bidragande andelen hårdgjord yta vara större än i till exempel ett skogsområde. Därför klassificeras ofta områden med avseende på bebyggelsetyp och avrinningen inom området beräknas med hjälp av erfarenhetsvärden.

Eftersom att olika typer av ytor ger upphov till olika flöden blir det nödvändigt att ta fram dessa avrinningskoefficienter genom att göra områdesklassificeringar. Vanligtvis kan områden klassificeras genom manuellelt utifrån flygfoton, laserdata och kartor över ledningsnät och fastigheter. Dock kan flera tusen fastigheter eller områden ingå i en och samma modell. Därför blir arbetet med klassificering så tidskrävande. En del områden kan även vara enklare att klassificera än andra. Ett större område kan till exempel bestå av endast villor eller flerfamiljshus, varpå det blir enklare att utgå från erfarenhetsvärden för avrinningskoefficienter. Det blir dock svårare när bebyggelsen är av varierande typ, och utspridd inom ett och samma avrinningsområde. Denna typ av klassificering eller gruppering av områden har använts länge och i olika syften. Dessa klassificeringssystem tenderar dock att inte utgöras av fastställda definitioner och resultaten utvärderas oftast endast kvalitativt och inte kvantitativt. Därför skulle framtagandet av en ny arbetsmetod betydligt minska påverkan av den mänsliga faktorn. Detta examensarbetet ämnar undersöka huruvida det är möjligt att utveckla alternativa och mer tidseffektiva metoder för områdesklassificering, samt att på ett kvantitativt sätt utvärdera framtagna metoder.

1.1 SYFTE OCH MÅL

Syftet med examensarbetet är att undersöka huruvida det går att utveckla alternativa metoder för framtagande av områdesklassificeringar för avrinningskoefficienter med avseende på bebyggelsetyp, utifrån till exempel vanligt förekommande kartmaterial. Därmed ämnar examensarbetet att förenkla det ofta tidskrävande manuella arbete detta framtagande innebär. Om resultaten påvisar att de utvecklade alternativa metoderna är användbara, syftar även examensarbetet fastställa vilken av metoderna som är lämpligast för användning.

Examensarbetet ämnar även att belysa vilken roll områdesklassificering spelar vid projektering eller modelleringsarbete i relation till olika metoder som använts för avrinningsberäkningar. En mindre del av arbetet ämnar även tillägnas åt att kunna klassificera områden med avseende på anslutningsförhållanden. Det slutgiltiga målet är att genom en jämförelse mellan olika metoder för områdesklassificering gentemot områden med sedan tidigare fastställd klassificering kunna avgöra om en ny alternativ metod är lämplig och mindre krävande med avseende på datatillgång och nedlagd arbetstid. Examensarbetet kan sammanfattas med följande frågeställningar:

- Går det att utveckla fungerande alternativa metoder för områdesklassificering med avseende på bebyggelsetyp?
- Går det att med vanligt förekommande kartmaterial utveckla en metod som klassificerar områden med avseende på deras anslutningsförhållanden?
- Hur påverkar valet av metod klassificeringen av områden med avseende på bebyggelsetyp?
- Vilka förutsättningar med avseende på dataunderlag kräver de olika metoderna och vilka begränsningar kan dataunderlaget innebära?
- Hur tidskrävande är arbetet för de olika metoderna?

1.2 AVGRÄNSNINGAR

För att motsvara ett examensarbete på 30 högskolepoäng under cirka 20 veckor har examensarbetet avgränsats till att utveckla och jämföra fyra olika metoder för områdesklassificering. Dessa omfattar tre metoder för områdesklassificering med avseende på bebyggelsetyp och en med avseende på anslutningsförhållanden. Examensarbetet behandlar ej avgränsningsmetoder av avrinningsområden, utan fokuserar primärt på klassificeringsmetoder. Därför används de områdesavgränsingar som är sedan tidigare definierade i dataunderlaget.

2 TEORI

Teoriavsnittet behandlar hydraulisk modellering som ligger till grund för examensarbetet, samt mycket av den teori som är nödvändig för att förstå hur de olika klassificeringsmetoderna utvecklades. Rektangulärklassificering och klassificering baserad på största sannolikhet användes i försök att områdesklassificera utifrån parameterskillnader, medan närmaste granne-algoritmen utgör grunden för hur en metod utvecklades i syfte att områdesklassificera baserat på intilliggande områden. Teoriavsnittet behandlar även de statistiska jämförelsemetoder som används för att utvärdera och tolka metodernas prestanda och resultat.

2.1 HYDRAULISK MODELLERING OCH AVLOPPSSYSTEM

Hydraulisk modellering avser modellering som omfattar fluiders dynamik. Detta innebär modellering av vätskors rörelse, och skulle till exempel kunna vara modelleringen av ett spillvattensystem. En modell motsvarar en förenklad bild av verkligheten. Hydraulisk modellering bör inte förväxlas med hydrologisk modellering som utgör ett något bredare begrepp. Hydraulik, eller tillämpad hydromekanik är snarare en gren inom hydrologi (Nationalencyklopedin, 2019c). Avloppsvatten avser de system som avleder vatten från hushåll, industrier och andra verksamheter, samt smält eller framträngt vatten och nederbörd. Detta för att skydda bebyggelse från översvämningsrisk och upprätthålla sanitära förhållanden (Nationalencyklopedin, 2019a). Detta omfattar alltså spill-, drän- och dagvatten. Spillvatten är det förbrukade vatten som avleds från fastigheter i form av till exempel toalett-, disk- och badvatten. Dränvatten kan istället komma från husgrunder och annan markavvattning. Dagvatten är det vatten som tillfälligt rinner på markytan eller konstruktioner. Detta omfattar inte endast regn, utan kan även utgöras av till exempel smältvatten från is och snö eller grundvatten som har framträngt vid markytan (Nationalencyklopedin, 2019b).

Vid modellering av vatten- och avloppssystem är det viktigt att ta hänsyn till dagvattenflödet inom avrinningsområdet, då detta oftast blir det största och dimensionerande flödet för många ledningar (Svenskt Vatten, 2010). Avrinningsområdet definieras som det uppströms belägna område där vatten dräneras till en bestämd punkt (SMHI, 2018). Områdets gränser utgörs av vattendelare mot andra avrinningsområden, alltså höjdryggar som delar de olika vattenflöden åt olika håll. Dessa vattendelare kan oftast bestämmas med hjälp av topografin inom området, då vatten rinner inom olika avrinningsområden beroende på vilken sida av vattendelaren som betraktas. I stort sett så finns det inga system som inte påverkas av dagvatten. Även de system som är speciellt avsedda för spillvatten har normalt någon form av dagvattenpåverkan. Detta kan ha ett flertal orsaker. Ofta beror det på att överläckage från dagvattennätet eller otäta brunnslock, men det kan även handla om till exempel bidragande avrinning från takytor som felaktigt har kopplats till systemet.

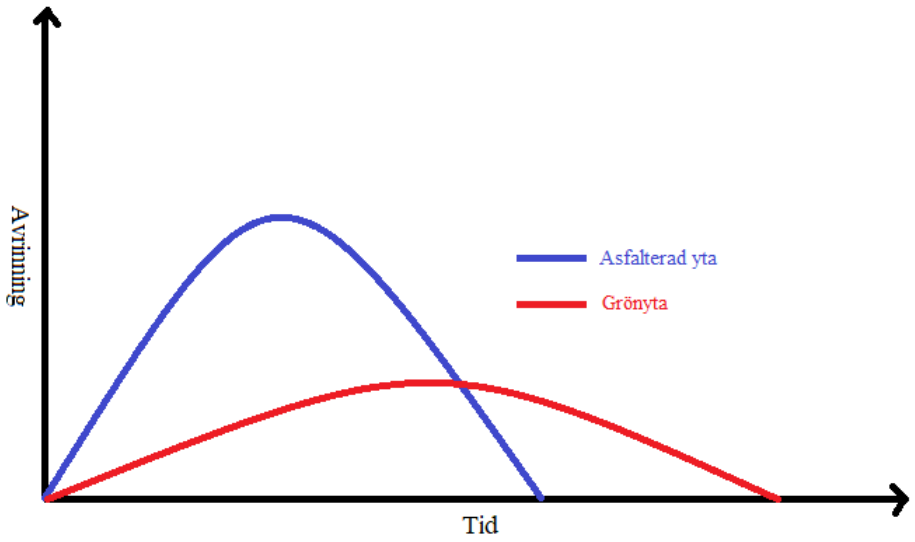
Hydrauliska modeller används i dagsläget som underlag för beslutsfattande inom ett flertal områden (Blomquist m. fl., 2016). Ofta används dessa till exempel som underlag för anpassning efter klimat och skyfall. Med hjälp av dessa typer av modeller kan effekten av

olika åtgärder utvärderas då en mängd olika utfall kan simuleras. Därmed blir själva modelleringen en viktig del av arbetet i planeringsfasen när områden ska exploateras, byggas ut, eller vid projekt som omfattar vatten- och avloppssystem.

När det genom modellering går att optimera för kapacitet och prioritering av åtgärder kan resurser används där de är som mest effektiva. När det kommer till att sätta upp en modell finns det väldigt många olika tillvägagångssätt. Därför är det viktigt att modellörer kan lägga sig på rätt nivå intitialt och arbeta resultatinriktat. När det kommer till att modellera eller beräkna flöden från dag- eller spillvattensystem kan själva modelleringsarbetet vara väldigt omfattande beroende på vilken detaljnivå som efterfrågas. Detta innebär att valet av modelleringsmetod orsakar en viss svårighet vid bedömning av analysresultatens kvalitet (Blomquist m. fl., 2016). Beroende på underlagskvalitetet av den tillgängliga datan kan resulterande flödesberäkning vara av varierande noggrannhet (Arnell, 1980). Dataunderlaget är oftast tillgängligt genom genom modellbeställare, databaser eller myndigheter som till exempel lantmäteriet (Blomquist m. fl., 2016). Geografiska informationssystem och analysverktyg används sedan för att behandla datan. En del metoder är dock endast tänkta att ge approximativa värden. Noggranhetskravet blir därmed en avvägning mellan vilken detaljnivå som är nödvändig för uppdraget och om eventuella kostnadsbesparingar kan göras genom att använda en mindre tidskrävande approximativ metod.

2.1.1 Flöden i ett avrinningsområde

De flöden som uppstår inom ett avrinningsområde delas ofta in i två olika kategorier. Dessa är direkt nederbördspåverkan och indirekt nederbördspåverkan (Blomquist m. fl., 2016). Direkt nederbördspåverkan avser de flöden som orsakas av nederbörd direkt på de hårdgjorda avrinningsbidragande ytor som är ansluta till avloppsledningsnätet i fråga. Den indirekta påverkan avser övriga flöden som belastar nätverket vid nederbörd. Dessa kan bestå av till exempel överläckage eller en snabb bildning av grundvatten som har dränerats till systemet avsett för spillvatten. Avrinningen och de flöden som uppstår inom området beror i sin tur på ett flertal olika saker. Dessa flöden påverkas av bland annat exploateringsgrad, andelen hårdgjord yta inom området, anslutningsförhållanden, lutning och nederbördsintensitet (Svenskt Vatten, 2010). Därmed blir det viktigt att ta hänsyn till att olika typer av ytor ger upphov till olika typer av flöden. Hårdlagda och sammanhållna ytor ger upphov till större avrinning under en kortare tid, med ett intesivare händelseförlopp. Andra ytor, som grönytor, åkermark eller övrig mark med mer växtlighet, har istället en buffrande effekt på händelseförloppet. Det resulterar i mindre avrinningstoppar och ett långsammare händelseförlopp. Detta illustreras i hydrografen i Figur 1.



Figur 1. *Hydrograf som teoretiskt visar på skillnaden i den avrinning som uppstår på en hårdlagd yta som asfalt och en grönlagd yta med växtlighet.*

Därmed kommer ett bebyggt område ge upphov till betydligt större avrinning under kortare tid i jämförelse med ett obebyggt eller icke exploaterat område. För att kunna modellera dessa skillnader före och efter urbanisering så kan GIS-verktyg vara mycket användbart för att inhämta data, då kunskap om ett områdes ytkaraktär möjliggör tillräcklig dimensionering av ledningssystemen i fråga.

2.1.2 Vanliga beräkningsmetoder för bidragande andel yta

För att uttrycka hur mycket av nederbörden som avrinner efter förluster som infiltration och absorption av växtlighet, magasinering i markytan eller evapotranspiration används en avrinningskoefficient (φ). Detta är ett tal mellan noll och ett som uttrycker hur stor andelen avrinnande vatten är (Svenskt Vatten, 2010). Det finns därmed ett flertal sätt att bestämma dagvattenflödet genom att bestämma andelen hårdlagd yta inom ett avrinningsområde. I Sverige används huvudsakligen tre olika sätt (Blomquist m. fl., 2016). Dessa går ut på att den bidragande andelen antingen bestäms med hjälp av kalibreringar mot flödesmätningar, områdets bebyggelsetyp, eller typen av yta inom området.

Om flödesmätningar finns att tillgå, kan den bidragande arean som är ansluten bestämmas genom kalibrering (Metod 1). Detta är ett bra alternativ om den ledningsnätet inom avrinningsområdet är enhetligt. Om det inte är det, så är i verkligheten metodens areafördelning ojämnt distribuerad över området. Om det istället finns uppgifter om bebyggelsetyp, men inte data för typer av ytor inom området så kan arean bestämmas med bebyggelsetyper (Metod 2). Detta är ofta fallet när ett område ska bebyggas men helt färdigställda byggplaner saknas, och utformningen av området är ännu ej känd. Finns det däremot tillräckligt underlag om olika typer av ytor inom området, samt dess storlek, så kan den bidragande andelen bestämmas med hjälp av dessa (Metod 3). Då används en sammanvägd av-

rinningskoefficient för hela avrinningsområdet. Den beräknas på följande sätt (Svenskt Vatten, 2010):

$$\varphi = \frac{A_1 * \varphi_1 + A_2 * \varphi_2 + ...A_n * \varphi_n}{A_1 + A_2 + ...A_n} \tag{1}$$

där $A_1...A_n$ är de olika areorna inom området, vilka har sina motsvarande avrinningskoefficienter $\varphi_1...\varphi_n$. Summan av $A_1...A_n$ blir därmed densamma som den totala arean av hela avrinningsområdet. Avrinningskoefficienterna för de olika typerna av ytor hämtas ofta i form av litteraturvärden från tabeller. Dessa finns i Tabell 1. Beroende på tillgängligt underlag och huruvida modellen ska kalibreras eller inte, så finns det utrymme för vissa parameterförändringar.

Tabell 1. *Avrinningskoefficienter för olika typer av ytor dimensionerade för kortvariga regn (Svenskt Vatten, 2010).*

Typ av yta	**Avrinnings-koefficient,** φ
Tak utan ytmagasin	0,9
Betong- och asfaltyta, berg i dagen i stark lutning	0,8
Stensatt yta med grusfogar	0,7
Grusväg, starkt lutande bergigt parkområde utan nämnvärd vegetation	0,4
Berg i dagen i inte alltför stark lutning	0,3
Grusplan och grusad gång, obebyggd kvartersmark	0,2
Park med rik vegetation samt kuperad bergig skogsmark	0,1
Odlad mark, gräsyta ängsmark m.m.	0–0,1
Flack tätbevuxen skogsmark	0–0,1

Då det är önskvärt att arbeta med en viss säkerhetsmarginal, resulterar metoden ofta i en överdimensionering (Blomquist m. fl., 2016). Om modellen ska kalibreras mot uppmätta flöden, bör avrinningsparametrarna istället bestämmas utifrån erfarenhetsvärden för avrinningsparametrar för olika typer av ytor. Dessa finns i Tabell 2.

Tabell 2. *Erfarenhetsvärden av avrinningskoefficienter för olika ytor (Larsson, 2010). A: Innerstad/Höghus förort, B: Villaförort/Villaområde i mindre ort, C: Övriga områden.* ***S****: Separerade ledningssystem,* ***K****: Kombinerat system med endast spillvattenledning i gatan,* ***D****: Delvis separerade system.*

	A			B			C		
Kategori	**S**	**K**	**D**	**S**	**K**	**D**	**S**	**K**	**D**
Tak: villa, fritidshus	0,90	0,90	0,09	0,45	0,45	0,09	0,32	0,27	0,06
Tak: flerfamiljshus	0,90	0,90	0,18	0,90	0,90	0,18	0,63	0,54	0,12
Tak: industri, övrigt	0,90	0,90	0,18	0,90	0,90	0,18	0,63	0,54	0,12
Hårdgjord yta	0,90	0,40	0,00	0,80	0,40	0,00	0,56	0,24	0,00
Grusad yta	0,20	0,10	0,00	0,20	0,10	0,00	0,14	0,06	0,00
Gatuyta	0,90	0,00	0,00	0,80	0,00	0,00	0,56	0,00	0,00

2.1.3 Områdesklassificering med avseende på bebyggelsetyp och anslutningsförhållanden

När det inte finns tillgänglig data för att bestämma avrinningskoefficienter utifrån andel bidragande hårdlagd yta inom avrinningsområdet, så kan dessa bestämmas genom att avgöra vilken bebyggelsetyp som är dominerande inom avrinningsområdet (Blomquist m. fl., 2016). Då används en avrinningskoefficient för hela avrinningsområdet eftersom att det inte går att göra en detaljerad beräkning utan tillräckligt underlag. Typiska värden för avrinningskoefficienter baserade på bebyggelsetyp finns i Tabell 3.

Tabell 3. *Sammanvägda avrinningskoefficienter för olika bebyggelsetyper dimensionerade för kortvariga regn (Svenskt Vatten, 2010).*

Bebyggelsetyp	**Flackt**	**Kuperat**
Slutet byggnadssätt, ingen vegetation	0,70	0,90
Slutet byggnadssätt med planterade gårdar	0,50	0,70
Öppet byggnadssätt (flerfamiljshus)	0,40	0,60
Radhus, kedjehus	0,40	0,60
Villor, tomter $< 1000\ m^2$	0,35	0,45
Villor, tomter $> 1000\ m^2$	0,20	0,30

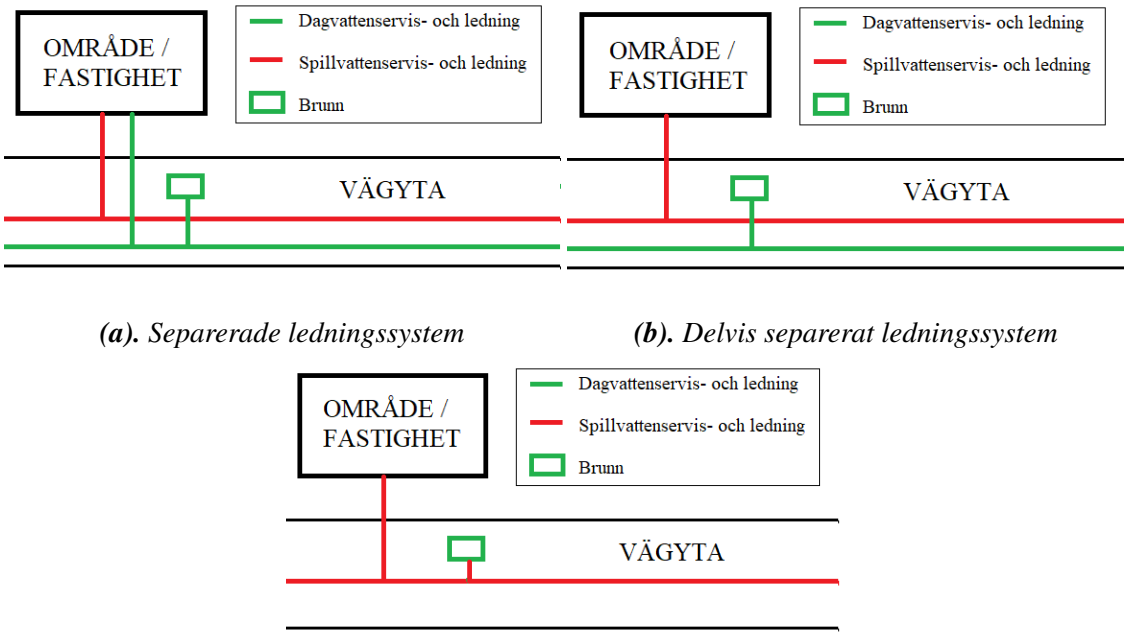
Erfarenheter visar på att värdena i Tabell 3 ofta resulterar i en överdimensionering av ledningsnät (Blomquist m. fl., 2016). Detta beror på att en viss säkerhetsmarginal är nödvändig när bebyggelsens utformning fortfarande är okänd. Även i denna metod är det då bättre att utgå från erfarenhetsvärden. Dessa finns i Tabell 4.

Tabell 4. *Erfarenhetsvärden av avrinningskoefficienter för olika bebyggelsetyper (Larsson, 2010). A: Innerstad/Höghus förort, B: Övriga områden.* ***S****: Separerade ledningssystem,* ***K****: Kombinerade system med endast spillvattenledning i gatan,* ***D****: Delvis separerade system.*

	A			B		
Kategori	**S**	**K**	**D**	**S**	**K**	**D**
Slutet byggnadssätt, ingen vegetation	0,74	0,33	0,052	0,28	0,21	0,028
Slutet byggnadssätt med planterade gårdar	0,53	0,24	0,037	0,2	0,15	0,02
Öppet byggnadssätt (flerfamiljshus)	0,42	0,19	0,03	0,16	0,12	0,016
Radhus, kedjehus	0,42	0,19	0,03	0,16	0,12	0,016
Villor, tomter $< 1000\ m^2$	0,26	0,12	0,019	0,1	0,075	0,01
Villor, tomter $> 1000\ m^2$	0,16	0,072	0,011	0,06	0,045	0,006
Park, gräsyta med mera	0,053	0	0	0,02	0	0

Begreppen i Tabell 2 och 4 som berör områdenas anslutningsförhållanden kräver dock en förklaring. Enligt Hammarlund (2019) har de återgivits fel i Svenskt Vattens publikation av Blomquist m. fl. (2016). Därför har förändringar gjorts för denna rapport, och begreppen så som de används i rapporten förklaras i detalj nedan.

Beteckningen *S* syftar på separerade anslutningsförhållanden där det finns både spill- och dagvattenledningar vilka är anslutna via serviser till både fastighet och dagvattenbrunnar. Därmed finns det separata serviser för spill- och dagvatten. Det är så områden ofta utformas idag eftersom det finns praktiska skäl för att skilja på spill- och dagvatten (Blomquist m. fl., 2016). *D* syftar på delvis kombinerade system där det finns både spill- och dagvattenledningar, men där endast vägytan är ansluten till dagvattenledningsnätverket. Fastigheternas dagvatten är istället anslutet till spillvattenledningsnätverket. Detta är vanligast förekommande i områden där det endast har funnits ett ledningsnätverk för spillvatten, där dagvattenledningar har dragits i efterhand. Då är det vanligt att inte kräva att fastigheterna separerar sina ledningsnät eftersom att det ofta är både komplicerat och kostsamt när det handlar om äldre fastigheter. Kostnaden faller ofta även då på privatpersoner. *K* syftar i sin tur på kombinerade ledningssystem. Dessa härstammar från när endast en ledning användes till både spill- och dagvatten. Skillnaderna illustreras i Figur 2.



(a). *Separerade ledningssystem* ***(b).*** *Delvis separerat ledningssystem*

(c). *Kombinerat ledningssystem*

Figur 2. *Olika anslutningsförhållanden*

Bortsett från dessa klassificeringar med avseende på anslutningsförhållanden finns det ytterligare alternativ (Hammarlund, 2019). Det kan till exempel gå diken eller vattendrag intill fastigheten. Ibland saknas ett ledningsnät helt och hållet, samt även diken eller vattendrag. Enklast är att kategorisera områden utefter de typer av anslutningsförhållanden som visas i Figur 2. Vanligast är troligen att kommuner klassificerar områden som områden med antingen separerade eller kombinerade system. Sammanfattningsvis skulle områden däremot kunna kategoriseras med avseende på anslutningsförhållanden på följande sätt:

- Separerat (S)
- Delvis kombinerat (D)
- Kombinerat (K)
- Om dag- och spillvattenserviser finns till fastigheten och endast dagvattenledning finns i gatan
- Om endast spillvattenservis finns till fastigheten och endast dagvattenledning finns i gatan
- Om diken och/eller vattendrag finns intill fastigheten
- Om serviser saknas till fastigheten
- Om ledningsnät i gatan, samt dike och/eller vattendrag saknas.

Att använda sammanvägda avrinningskoefficienter för olika bebyggelsetyper fungerar bra med tanke på att den bygger på överslagsberäkning (Larsson, 2010). Framförallt då metoden är mycket mindre tidskrävande än många andra metoder, då den inte nödvändigtvis förutsätter någon manuell kartering. Vidare bedöms metoden fungera likvärdigt väl för alla olika typer av områden. I mer tätbebyggda områden kan däremot lutningen bli en mer betydande faktor för avrinningskoefficienten, medan effekten av lutningen verkar avta ju större avrinningsområdet är. Enligt Larsson (2010) blir resultatet mer tillförlitligt om avrinningskoefficienterna kategoriseras beroende på vilken typ av ledningsnät och anslutningsförhållanden som finns i området, i jämförelse med att endast utgå från bebyggelsetyp. Därmed är det ofta även nödvändigt att bestämma områdets anslutningsförhållanden.

2.2 BILDKLASSIFICERING

Det har gjorts olika försök att klassa avrinningsområden med avseende på ett flertal olika parametrar, i olika syften. Däremot saknas ett generellt ramverk för precis hur det ska göras (Sivapalan m. fl., 2007). Det är däremot fastställt att det går att erhålla approximativa flöden med enbart kunskap om avrinningsområdens strukturella karaktär och bebyggelse (Larsson, 2010).

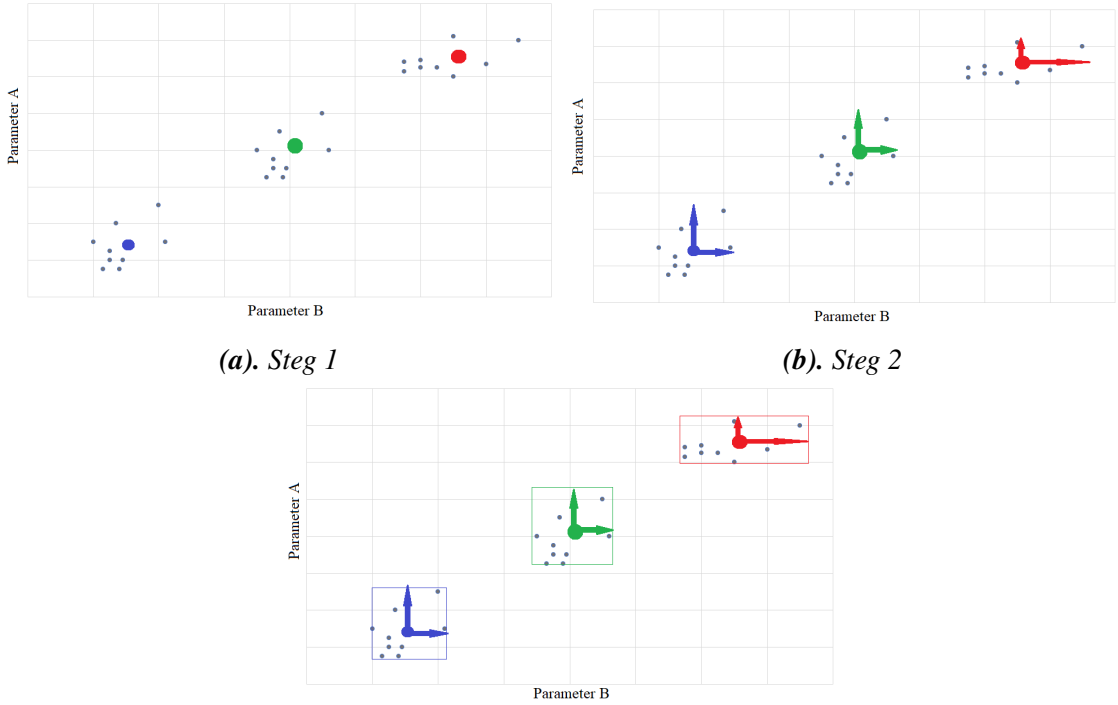
Bildklassifikation används ofta för att klassdefiniera landområden, dock oftast inte endast med avseende på bebyggelse utan snarare områdestyper som åker, vatten och skog. Detta går ut på att analysera färgerna i olika rasterband för att på så sätt kunna klassificera alla olika pixlar i en bild (ESRI, 2019c). Oftast är dessa bilder till exempel flyg- eller satellitbilder. Beroende på hur involverad modelleraren är i klassificeringsprocessen så finns det två huvudsakliga klassificeringsmetoder inom bildklassifikation - övervakad och icke övervakad klassificering. Icke övervakad klassificering bygger på att en dator definierar ett antal klasser och fattar beslut utan processövervakning baserat på pixlarnas spektrala egenskaper. Övervakad klassificering bygger på att datorn fattar beslut baserat på vilka klasser som har assignerats en viss andel av pixlarna i form av träningsdata, och därmed är det användaren som definierar tillgängliga klasser. Det är svårt att avgöra vilken typ av klassificeringsmetod som bör användas, då metodvalet måste anpassas efter syftet med klassificering och underlagsdatat (Perumal och Bhaskaran, 2010). Därmed går det inte att säga vilken klassificeringsmetod som är bäst av de metoder som finns tillgängliga. Dock är rektangulärklassificering (Parallelepiped Classification) och klassificering baserad på största sannolikhet (Maximum Likelihood Classification) två av de vanligaste metodvalen inom övervakad klassificering.

2.2.1 Rektangulärklassificering

Rektangulärklassificering bygger på en relativt enkel matematisk algoritm och passar bra för dataunderlaget där olika klasser skiljer sig avsevärt utan överlappning (Ukrainski, 2017). Algoritmen består av tre steg. I det första steget beräknas centrum för varje klass i den spektrala enhetsrymden. När det handlar om bildklassifikation används oftast bandens medelvärde för ljusstyrka och dess standardavvikelser. Därefter i steg två definieras

extrempunkter för varje klass i form av minimum och maximumvärden . I det sista tredje steget så definierar algoritmen linjer som formar rektanglar utifrån extremvärden. Dessa tre steg illustreras i Figur 3 för två godtyckliga parametrar. Detta kan dock även tillämpas för flera parametrar i flera dimensioner.

Denna typ av klassificering fungerar som nämnt dock sämre när dataunderlaget överlappar. Då förblir objekt eller pixlar icke klassificerade, eller erhåller ofta en felaktig klassificering. Därför fungerar algoritmen som bäst när klasser utgör separata kluster, så som i Figur 3 där tre exempel visas.



(a). *Steg 1*

(b). *Steg 2*

(c). *Steg 3*

Figur 3. *De tre stegen som utgör algoritmen för rektangulärklassificering för två parametrar (A och B) och tre olika klasser (röd, grön och blå).*

2.2.2 Klassificering baserad på största sannolikhet

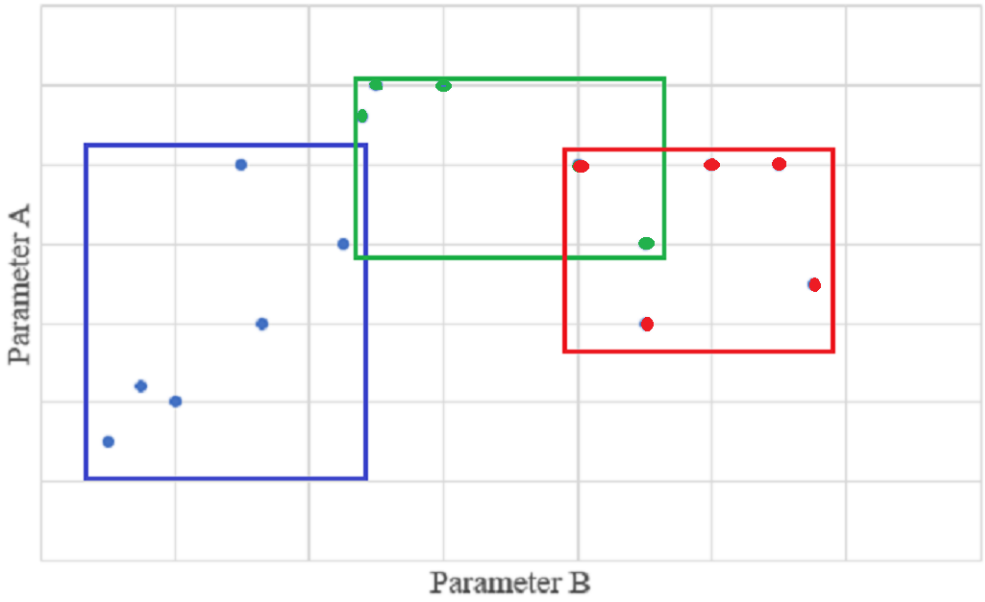
Bland övervakade klassificeringsmetoder är klassificering baserad på största sannolikhet bland de absolut vanligast förekommande, framförallt då denna metod har visat sig prestera väl för ett flertal olika syften och system (Bolstad och Lillesand, 1991). En nackdel med metoden är dock att den involverar mycket beräkning, vilket kan kräva mycket processorkraft för större områden eller bilder med hög upplösning. Den förutsätter även att dataunderlaget är normalfördelat, och därför är det rekommenderat att logaritmera dataunderlaget om detta ej är fallet (Ahmad och Quegan, 2012). Denna metod bygger på att välja ut träningsdata för olika klasser, och sedan låta datorn utföra resterande klassificeringar på egen hand, för att därefter granska dem. Till skillnad från att endast utgå från extremvärden, används istället medelvärden och kovariansen för de olika parametrarna för

att klassificera baserat på största sannolikhet (ESRI, 2019a). Detta är härlett från Bayes sats som lyder (Ahmad och Quegan, 2012):

$$P(i \mid \omega) = \frac{P(\omega \mid i)P(i)}{P(\omega) = \sum_{i=1}^{M} P(\omega \mid i)P(i)} \tag{2}$$

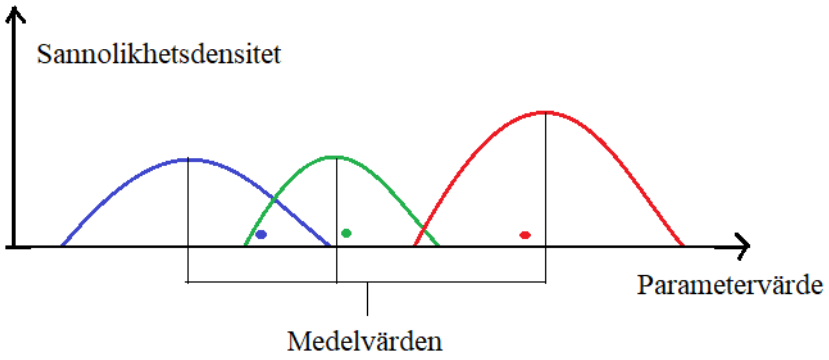
där $P(\omega \mid i)$ är sannolikhetsfunktion, $P(i)$ är prioritetskonstant, och $P(\omega)$ är sannolikheten att ω observeras i M antal klasser. Prioritetskonstanten utgörs av sannolikheten att att klassen i fråga påträffas i det studerade området.

Sannolikheten för varje enskild pixel eller varje enskilt objekt beräknas därmed för vardera klass, och den med högst sannolikhet att tillhöra en särskild klass är den som väljes. Detta tillämpas när data ser ut som i Figur 4, som illustrerar ett exempel där med för tre olika klasser och två parametrar (därmed är detta exempel tvådimensionellt).



Figur 4. *Överlappande data tillhörande tre olika klasser (röd, grön och blå) för två parametrar.*

Klassificeringar genomföres därmed genom att bestämma hur stor sannolikheten är för att en särskild punkt, pixel eller ett objekt, tillhör en särskild klass. Detta illustreras med ett exempel på sannolikhetsdensitetsfunktioner eller täthetsfunktioner för tre olika punkter och tre olika klasser i Figur 5. En täthetsfunktion är en kontinuerlig stokastisk variabel av en funktion, vilken motsvarar lutningen av variabelns fördelningsfunktion (Nationalencyklopedin, 2019d). Därför motsvarar arean mellan två gränser under täthetsfunktionen sannolikheten att att variabeln antar ett värde mellan de två gränserna.



Figur 5. *Överlappande sannolikhetsdensitetsfunktioner för tre olika klasser (röd, grön och blå) med tre färgklassificerade punkter baserade på största sannolikhet.*

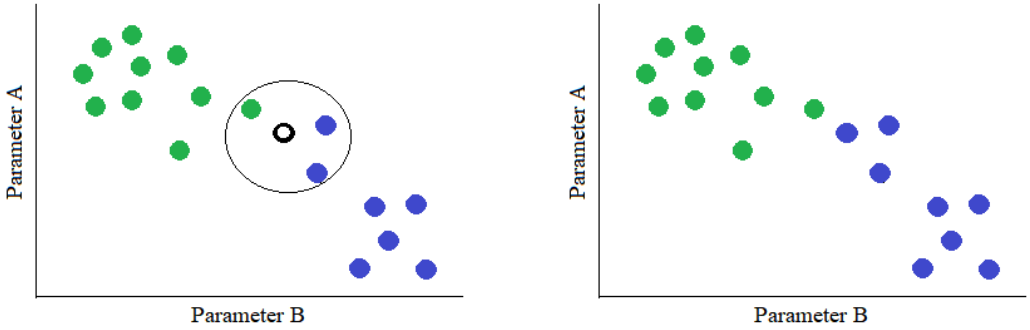
2.3 KLASSIFICERING BASERAD PÅ INTILLIGGANDE OMRÅDEN

Redan år 1970 hänvisade Waldo R. Tobler till vad han kallade geografins första lag, som därefter även har kallats för Toblers lag. Denna lag dikterar att allting är relaterat till allting annat, men att näraliggande enheter relaterar mer till varandra i jämförelse med andra enheter som ligger längre bort (Tobler, 1970). Detta går även att tillämpa på de områden som avses klassificeras. Framförallt eftersom att många bebyggelsetyper tenderar att samlas i kluster. I de flesta städer planeras ellet uppstår till exempel industri- och villaområden separat, och det är ovanligare att hitta den ena bebyggelsetypen blandad med den andra.

2.3.1 Närmaste granne-algoritmen

Den närmaste granne-algoritmen (K-Nearest Neighbour Algorithm eller KNN-algoritmen) har varit väldigt vanligt förekommande när det kommer till att klassificera flera olika typer av data ända sedan den presenterades av Cover och Hart år 1967 (Kataria och Singh, 2013). Algoritmen bygger på att identifiera näraliggande grannar och sedan klassificera data baserat på grannarnas egenskaper. Detta kan göras genom att ge algoritmen träningsdata, vilken används för att utföra alla klassificeringar (Kataria och Singh, 2013).

Däremot är det viktigt att skilja på huruvida det endast är den närmaste grannen som används för att klassificera en datapunkt, eller om det är flera grannar. Det är där faktorn K i K-Nearest Neighbour Algorithm kommer in i bilden och gör metoden enkel. Faktorn K bestämmer just hur många grannar det är som används för att utföra klassificeringen (Kataria och Singh, 1967). Metoden är enkel för att den endast använder majoriteten av intilliggande punkter för att utföra klassificeringen. Algoritmen har visats vara väldigt effektiv för att uppnå hög noggrannhet som klassificeringsverktyg. Däremot är algoritmen ofta väldigt beräkningstung eftersom att den ofta behöver beräkna avstånd mellan många datapunkter beroende på storleken på K (Hwang och Wen, 1998). I Figur 6 har K valts till att omfatta tre grannar. Eftersom att de blå punkterna är i majoritet av de tre grannarna, kommer den svartvita punkten att klassificeras som blå.



(a). *K väljes till 3 för att klassificera den svart-vita datapunkten.*

(b). *Den svartvita punkten har klassificerats som blå.*

Figur 6. *Hur en datapunkt klassificeras med hjälp av närmaste granne-algoritmen.*

2.4 STATISTISKA JÄMFÖRELSER

2.4.1 Utvärdering med förväxlingsmatriser

En förväxlingsmatris används ofta för att undersöka hur väl ett system eller en modell för klassificering presterar. Matrisen är tvådimensionell och indelad i kategorier längs varje sida. Den ena sidan motsvarar den egentliga klassen, medan den andra sidan motsvarar den klass som har blivit vald av systemet eller modellen (Ting, 2017). Denna metod används frekvent inom maskininlärning då den på ett enkelt sätt tillåter jämförandet av ett flertal olika modellers prestanda.

Förväxlingsmatrisens storlek bestäms av $n \times n$, där n är antalet klassificeringar som jämföres (Visa m. fl., 2011). I Figur 7 illustreras ett problem av andra klassen, när $n = 2$. Med antalet jämförelser växer därmed matrisen. I Figur 8 illustreras ett problem med fler olika alternativ. Det fungerar dock på exakt samma sätt som för ett problem av andra klassen likt det som illustrerades i Figur 7, och därför kan statistiska mått på prestanda erhållas för större modeller med fler klassificeringsalternativ.

	SANN POSITIV	SANN NEGATIV
UPPSKATTAD POSITIV	a	b
UPPSKATTAD NEGATIV	c	d

Figur 7. *Förväxlingsmatris av ett problem av andra klassen ($n = 2$).*

	SANN A	SANN B	SANN C	SANN ...
UPPSKATTAT A	a	b	c	...
UPPSKATTAT B	d	e	f	...
UPPSKATTAT C	g	h	i	...
UPPSKATTAT ...	...	...	...	...

Figur 8. *Förväxlingsmatris för ett problem av storleken* $n \times n$.

Förväxlingsmatrisen i Figur 7 kan användas för att illustrera hur en prestandamätning skulle kunna gå till. I denna förväxlingsmatris finns det två alternativ att gissa på för en modell. Ett positivt resultat, eller ett negativt resultat. Om ett egentligt resultat (det sanna resultatet) är positivt och en hypotetisk modell skulle uppskatta resultatet som positivt, så loggas detta i den gröna rutan betecknad med a. Om modellen istället skulle uppskatta resultatet som negativt när det egentligen är positivt så loggas detta i den röda rutan betecknad med c. Då det är känt hur många gånger en modell gissar fel, samt rätt, för varje enskild klass eller kategori, så är det möjligt att erhålla ett flertal statistiska mått på modellens prestanda. Med Figur 7 så bestäms noggrannheten på följande sätt. Detta motsvarar hur stor del av tiden modellen gissar rätt.

$$\textit{Noggrannhet} = \frac{a+d}{a+b+c+d} \tag{3}$$

På liknande sätt kan avvikelsen bestämmas genom följande formel. Detta motsvarar hur ofta modellen gissar fel.

$$\textit{Avvikelse} = \frac{b+c}{a+b+c+d} \tag{4}$$

Vidare, så går det även att beräkna ett mått på hur ofta modellen gissar på ett positivt resultat, när det egentligen är negativt, samt hur stor del av tiden modellen gissar på ett negativt resultat, när det egentligen är positivt. Dessa mått kan kallas för sanna och falska positiva och negativa gissningar. För att beräkna till exempel hur stor del av tiden modellen gissar på falska positiva resultat (som loggas i den röda rutan betecknad b), delas andelen av dessa gissningar med totala antalet gånger då resultatet egentligen har varit positivt. För en falskt positiv gissning erhålles därmed formeln:

$$\textit{Falskt positiv} = \frac{b}{b+d} \tag{5}$$

Med hjälp av dessa mått på hur stor del av tiden falska och negativa positiva och negativa gissningar inträffar, så är det även möjligt att bestämma precision samt "recall", här hänvisat till som känslighet, för de enskilda alternativen som finns tillgängliga. Precisionen blir ett mått på hur stor del av tiden som modellen har rätt när den gissar på ett enskilt alternativ. Känsligheten blir istället ett mått på hur stor del av tiden modellen till exempel valde rätt alternativ när det egentliga svaret var alternativ A. Med utgång i förväxlingsmatrisen i Figur 8 blir därmed formlerna för precision och känslighet följande för alternativ A.

$$\textit{Precision} = \frac{a}{a + b + c + ...} \tag{6}$$

$$\textit{Känslighet} = \frac{a}{a + d + g + ...} \tag{7}$$

För att bättre kunna utvärdera en metods prestation för olika klasser används ett F1-värde (Sokolova m. fl., 2006). Detta är ett harmoniskt medelvärde av precisionen och känsligheten. Detta straffar extremvärden och blir därför ett bättre mått för noggrannhet när prestandan för enskilda klasser jämföres. Detta varierar mellan noll och ett, där det råder jämn balans mellan precision och känslighet när värdet närmar sig ett. Formeln för F1-värdet är:

$$\textit{F1} = 2 * \frac{\textit{Precision} * \textit{Känslighet}}{\textit{Precision} + \textit{Känslighet}} \tag{8}$$

2.4.2 Cohens kappa

Cohens kappa (κ) är ett statistiskt mått som fungerar som ett alternativ till noggrannhet (Tallón-Ballesteros och Riquelme, 2014). Skillnaden är att det kompenserar för slumpen. Cohens kappa tar på samma sätt hänsyn till slumpen genom att i det statistiska måttet uttrycka hur en klassificeringsmetod utförde sin uppgift i jämförelse med om klassificeringarna hade valts slumpmässigt, och uttrycker därmed en skillnad i noggrannhet och avvikelse. Cohens kappa beräknas på följande sätt (Cohen, 1960).

$$\kappa = \frac{p_0 - p_c}{1 - p_c} = 1 - \frac{1 - p_0}{1 - p_c} \tag{9}$$

där p_0 är andelen gånger rätt klassificering valts och p_c andelen gånger rätt klassificering valts om den vanligaste klassificeringen hade valts varje gång. Alltså jämföres om slumpen hade kunnat uppnå samma resultat. I rapporten kommer p_c att hänvisas till som *nollratio*. Om nollration är lika med ett skulle det därmed innebära att den mest förekommande klassen alltid är korrekt klassificering. Om nollration istället var lika med noll skulle det innebära att den mest förekommande klassen aldrig utgjorde korrekt klassificering.

Måttet har gränsvärden. När $\kappa = 0$, så hade slumpen lika gärna kunnat avgöra och när $\kappa = 1$ så råder perfekt överensstämmelsegrad (Cohen, 1960). För att på ett konsistent sätt kunna tolka den relativa överensstämmelsegraden av olika κ-värden, kategoriserade Koch och Landis (1977) olika värden. Dessa går att se i Tabell 5. Intervallen är arbiträra, men utgör användbara riktmärken för att utvärdera överensstämmelsegrader.

Tabell 5. *Olika värden för Cohens κ och deras överensstämmelsegrad (Koch och Landis, 1977).*

κ-värde	**Överensstämmelse**
<0,00	Dålig
0,00 - 0,20	Måttligt dålig
0,21 - 0,40	Måttligt bra
0,41 - 0,60	Bra
0,61 - 0,80	Mycket bra
0,81 - 1,00	Nästintill perfekt

2.4.3 Shapiro-Wilks test

Om en metod förutsätter normalfördelad parameterdata, krävs underlag för hur fördelningen av parameterdatan förhåller sig till en normalfördelningskurva. För att ta reda på om dataunderlag är normalfördelat används ofta ett Shapiro-Wilk-test (Shapiro m. fl., 1968). Testet går ut på att en nollhypotes formuleras som hävdar att datan är normalfördelad, vilken sedan antingen kan styrkas eller förkastas. Ett flertal W-värden beräknas först enligt Ekvation 10.

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{10}$$

där x_i är mätvärden från provet ur populationen, $\bar{x}$ är medelvärdet av alla dessa mätvärden, och a_i är koefficienter som beräknas med hjälp av förväntade randomiserade värden ur mätvärdenas fördelning, samt mätvärdenas kovarians.

För att nollhypotesen ska förkastas behöver alla uppmätta värden ligga så nära noll som möjligt. Ett konfidensintervall beräknas för samtliga beräknade W-värden. Ofta väljes en signifikansnivå då $\alpha = 0,05$, och därmed beräknas 95%-konfidensintervall (Helsel och Hirsch, 2002). Detta innebär att om 95% av alla parametervärden ligger tillräckligt nära noll så kan den formulerade nollhypotesen förkastas, vilket motsvarar ett p-värde (probabilitetsvärde) mindre än 0,05. Detta hade inneburit att den testade parametern inte var normalfördelad. Det bör även nämnas att det även går att uppskatta huruvida en parameter är normalfördelad genom att undersöka var medianen ligger i ett lådagram, vilket kan vara hjälpsamt vid översiktlig datainspektion. Ligger den ungefär i mitten av lådagrammet är det troligt att dataunderlaget är normalfördelat.

2.5 MODELLPORTABILITET

Med portabilitet avses möjligheten att applicera en modell eller metod på mer än ett dataset. Detta kan ofta avse mjukvara i olika miljöer (TechTarget, 2019), men kan även appliceras på modeller eller klassificeringsmetoder. Alltså utgör modellens portabilitet dess förmåga att appliceras på fler ställen, utöver där modellen har utvecklats.

Många studier har gjorts i syfte att förbättra portabiliteten för modeller som används för att klassificera områden med avseende på landanvänding (Matasci m. fl., 2019). Målet med portabilitet kan definieras som att lyckas applicera en klassificeringsmodell med goda resultat på mer än ett dataset. Eftersom olika dataset ofta skiljer sig åt så kommer ofta resultatet för klassificeringsmodeller att variera beroende på var de appliceras. Att mäta portabilitetet blir dock mer problematiskt. För att kunna uppskatta portabiliteten för en modell måste den köras på så många olika dataset som möjligt. I en studie av Matasci m. fl. (2019) så kördes en övervakad bildklassificeringsmodell för landanvändning på 10 olika dataset och 100 olika genererade pixlar i två olika städer. F1-värden och Cohens κ-värden jämfördes sedan för de olika resultaten och klasserna. På så sätt gick det att diskutera hur modellen presterade på olika dataset. Därför utgör tillgängligt dataunderlag ibland en begränsning för hur väl en metod kan utvärderas.

3 MATERIAL OCH METOD

3.1 OMRÅDESBESKRIVNINGAR OCH DATAUNDERLAG

Ortofoton över Linköping samt Västervik med omnejd i upplösningen 0,25 meter erhölls från Lantmäteriet genom Sveriges lantbruksuniversitets databas (Lantmäteriet, 2019a). Dataunderlag i form av vektordata för olika former av kartmaterial erhölls från Linköpings och Västerviks kommun genom Tyréns. Dessa omfattade:

- Byggnadsytor
- Dagvattenserviser
- Dagvattenbrunnar
- Områden
- Spillvattenserviser
- Spillvattenbrunnar
- Vägkanter.

Dataunderlaget från Linköping omfattade även spill- och dagvattenledningar, vilket inte fanns med för Västervik. Vid inspektion av samtliga ledningsnät och serviser framgår det att underlaget ofta är otillräckligt. I dataunderlaget är det vanligt att serviser sällan når hela vägen fram till fastigheter samt att delar av ledningsnät saknas.

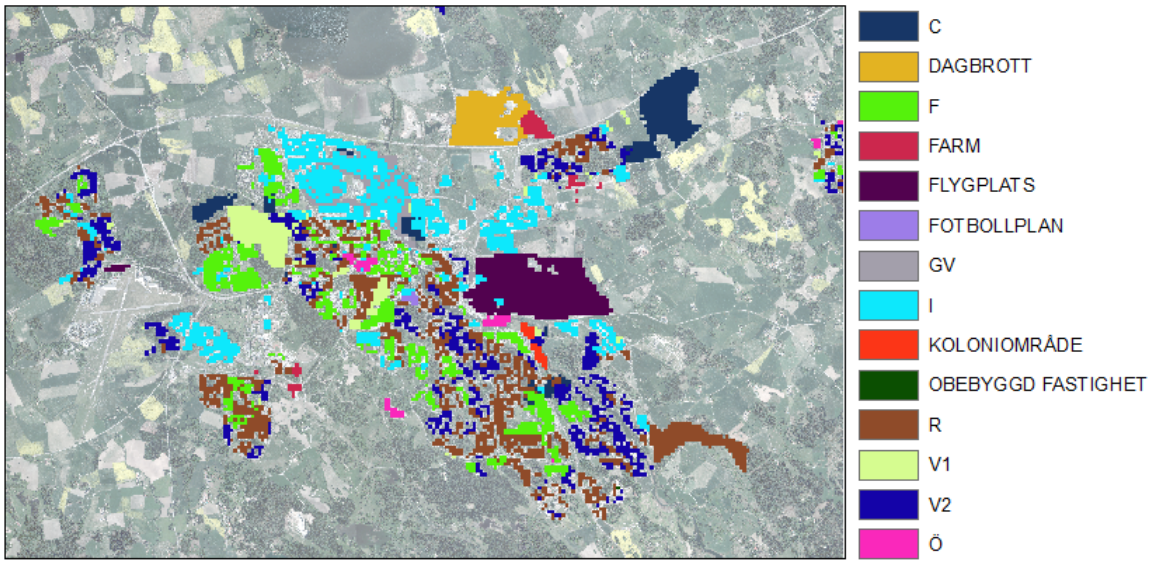
I Linköping var samtliga områden redan klassificerade sedan tidigare av personal från Tyréns med avseende på bebyggelsetyp. Inga områden i Västervik var klassificerade med avseende på bebyggelsetyp, men istället anslutningsförhållanden. Inga områden i Linköping var sedan tidigare klassificerade med avseende på anslutningsförhållanden.

3.1.1 Linköping

Alla områden var klassificerade med avseende på bebyggelsetyp av ett flertal anställda på Tyréns. Detta har gjorts över tid genom inspektion av varje område. I Figur 9 syns fördelningen av de olika bebyggelsetyperna. I Tabell 6 är det sammanställt hur många fastigheter det finns av varje bebyggelsetyp. Klassificeringen av bebyggelsetyp utgår från den vanliga typen av klassificering beskriven under Avsnitt 2.1.3, tillsammans med några ytterligare bebyggelsetyper som anställda på Tyréns har kompletterat med. Inga områden var klassificerade med avseende på anslutningsförhållanden.

Dataunderlaget inspekterades överskådligt innan användning och ett fåtal tydligt felaktiga klassificeringar gjordes om. Dessa felaktiga klassificeringar har troligtvis uppstått då flera fastigheter har klassificerats gruppvis eller av liknande anledning. Möjligheten finns dock

att enskilda felaktiga klassificeringar kvarstår, då det finns totalt 9002 redan klassificerade områden. Dessa områdesklassificeringar användes för att utvärdera klassificeringsmetoderna för bebyggelsetyp.



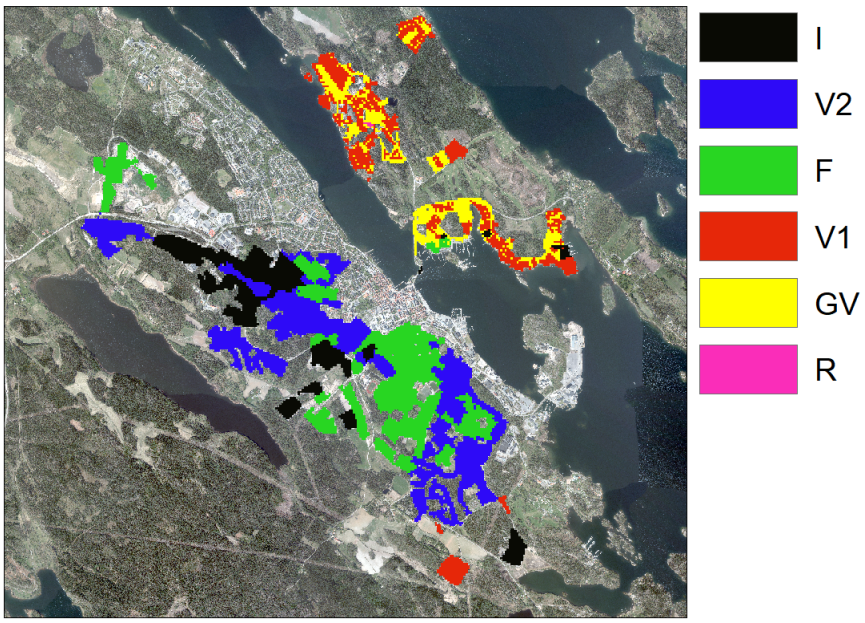
Figur 9. *Färgkarta över fördelningen av redan klassificerade områden i Linköping. ©Lantmäteriet. Bakgrundsbild: Ortofoto, 0,25 m färg (Lantmäteriet, 2019a).*

Tabell 6. *Förekomsten av olika redan klassificerade områden i Linköping.*

Bebyggelsetyp	**Antal**	**Beskrivning**
C	33	Tomtmark och liknande
DAGBROTT	3	Dagbrott
F	516	Områden bebyggda med flerfamiljshus
FARM	28	Farmer och gårdar
FLYGPLATS	2	Flygplatser
FOTBOLLPLAN	2	Fotbollsplaner
GV	15	Gatu- och vägyta
I	471	Industriområden
KOLONIOMRÅDE	2	Koloniområden
OBEBYGGD FASTIGHET	3	Obebyggda fastigheter
R	3226	Områden bebyggda med radhus
V1	47	Områden bebyggda med villor, tomter $< 1000\ m^2$
V2	4614	Områden bebyggda med villor, tomter $> 1000\ m^2$
Ö	40	Övriga områden

3.1.2 Västervik

Även i Västervik var bebyggelsetypen angiven som attribut för alla områden. Områdesfördelning går att se i Figur 10, samt i Tabell 7. Enstaka områden saknade klassificering, varför dessa sorterades bort. Ett fåtal områden var även klassificerade som *VÄXTHUS* och *V3*. Eftersom att det var okänt precis vad detta innebar, sorterades dessa områden bort och inkluderades inte i studien. Totalt fanns det 905 redan klassificerade områden.

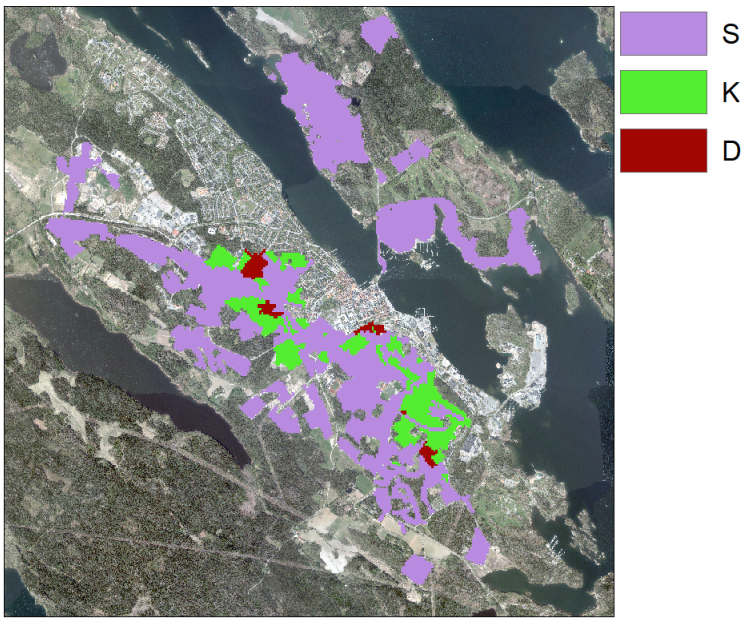


Figur 10. *Färgkarta över fördelningen av redan klassificerade områden i Västervik. ©Lantmäteriet. Bakgrundsbild: Ortofoto, 0,25 m färg (Lantmäteriet, 2019b).*

Tabell 7. *Förekomsten av olika redan klassificerade områden i Västervik.*

Bebyggelsetyp	**Antal**	**Beskrivning**
F	104	Områden bebyggda med flerfamiljshus
GV	73	Gatu- och vägyta
I	36	Industriområden
R	8	Områden bebyggda med radhus
V1	4	Områden bebyggda med villor, tomter $< 1000\ m^2$
V2	158	Områden bebyggda med villor, tomter $> 1000\ m^2$

Bortsett från bebyggelsetypen, var varje område klassificerat med avseende på anslutningsförhållanden. Det var angivet i attributform om varje område hade kombinerade spill och dagvattensystem, separerade system, eller delvis separerade system. Fördelningen av dessa totalt 926 områdena går att se i Figur 11 samt Tabell 8.



Figur 11. *Färgkarta över fördelningen av områdenas ledningssytem i Västervik, där D står för delvis separerat ledningssystem, S står för separerat ledningssystem och Kombinerat står för kombinerat ledningssystem. ©Lantmäteriet. Bakgrundsbild: Ortofoto, 0,25 m färg (Lantmäteriet, 2019b).*

Tabell 8. *Förekomsten av områden med olika ledningsnätsystem i Västervik.*

Ledningsnätsystem	**Antal**
(S) Separerade ledningssystem	799
(K) Kombinerade ledningssystem	99
(D) Delvis separerat ledningssystem	28

3.2 ANALYSVERKTYG

3.2.1 ArcGIS

Geografiska informationssystem (GIS) utgörs av program vilka används för samla in, hantera, och presentera olika geografiska data. ArcGIS är en plattform, eller ett programvarupaket som används för det ändamålet, att hantera och analysera geografisk data (ESRI, 2019b). Genom visualisering av data kan trender och mönster enklare analyseras och presenteras. ArcGIS utgör framförallt ett bra verktyg för att upptäcka mönster och kartlägga dessa med geografiskt dataunderlag. En programvarulicens erhölls av Uppsala universitet. Detta programvarupaketet användes eftersom att det är vanligt inom hydraulisk modellering, samt används frekvent av anställda på Tyréns. De verktyg som användes i ArcGIS beskrives i detalj under respektive metoddel.

3.2.2 R och RStudio

R är ett programmeringsspråk och en miljö som används för statistiska beräkningar och grafik (The R Project for Statistical Computing, 2019). Detta omfattar ett flertal statistiska tester, samt verktyg för klassificering, linjär och ickelinjär modellering. R är även skapat i öppen källkod, och finns därför tillgängligt utan kostnad under villkoren för Free Software Foundation's GNU General Public License i källkodsform (The R Project for Statistical Computing, 2019).

RStudio är ett grafiskt användarsnitt i mjukvaruform som har skapats för att användas med R, i syfte att göra verktygen mer lättillgängliga och användbara för en bredare målgrupp (RStudio, 2019). Även detta finns tillgängligt under samma villkor för öppen källkod.

Funktionerna **density** och **shapiro.test** användes för att skapa sannolikhetsdensitetsgrafer och för att kontrollera huruvida parameterdata var normalfördelad.

3.2.3 FME Desktop

Vid modelleringsarbetet och metodutvecklingen används främst FME Desktop utgivet av Safe Software ™, då detta är det program som vanligtvis används av Tyréns AB för dataintegration- och behandling. FME står för Feature Manipulation Engine och är en platform för dataintegration (Safe Software, 2019a). Programvaran bygger på att data av olika typer kan hanteras och manipuleras med hjälp av vad som kallas transformatorer, utan att någon kod behöver skrivas. Data som polygoner, punkter, linjer och liknanade hänvisas till som *Enheter*, och information tillhörande dessa som *Attribut*. Även detta är ett programvarupaket som består av bland annat FME Workbench och Data Inspector, vilka främst användes i analysarbetet. En programvarulicens erhölls av Uppsala universitet.

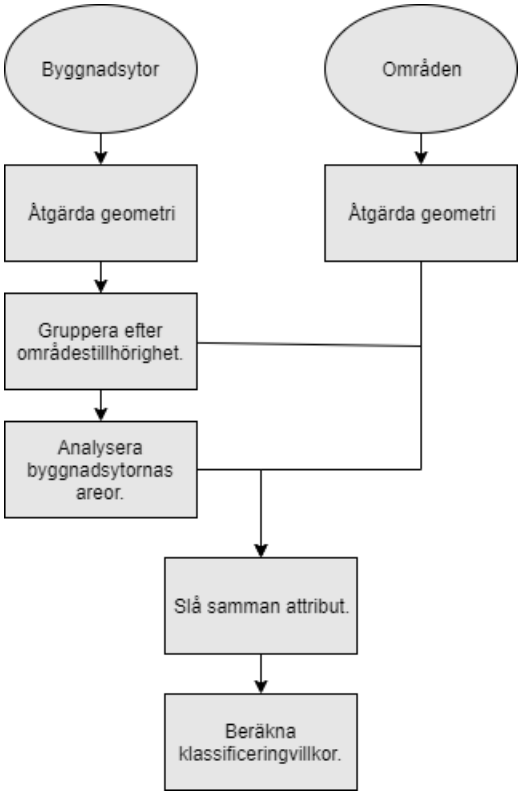
Eftersom att det främst var FME som användes i modelleringsarbetet finns samtliga transformatorer listade med sammanfattande beskrivningar i Bilaga A.

3.3 METOD I - REKTANGULÄRKLASSIFICERING OCH BYGGNADSYTOR

3.3.1 Klassificeringsvillkor

I syfte att utföra rektangulärklassificering på parametrar baserade på områdets area och deras byggnadsytor behövdes ett antal parametrar som klassificeringsvillkor. Dessa formulerades i form av minimum- och maximum-värden för varje tänkbar klassificering med avseende på bebyggelsetyp som redan var kända i underlaget från Linköping. Processen har gjorts överskådlig i ett flödesschema i Figur 12. Transformatorerna som användes är sammanställda med sina separata beskrivningar i Bilaga A. De olika parametrarna som klassificeringarna baserades på begränsades till:

- Antalet fastigheter inom ett område
- Arean för den minsta byggnadsytan inom ett område
- Arean för den största byggnadsytan inom ett område
- Kvoten mellan den största byggnadsytan inom ett område och områdets area
- Kvoten mellan den minsta byggnadsytan inom ett område och områdets area
- Medelvärdet av arean av byggnadytor inom ett område
- Summan av areorna av byggnadsytorna inom ett område



Figur 12. *Förenklat flödesschema över metoden för val av klassificeringsvillkor.*

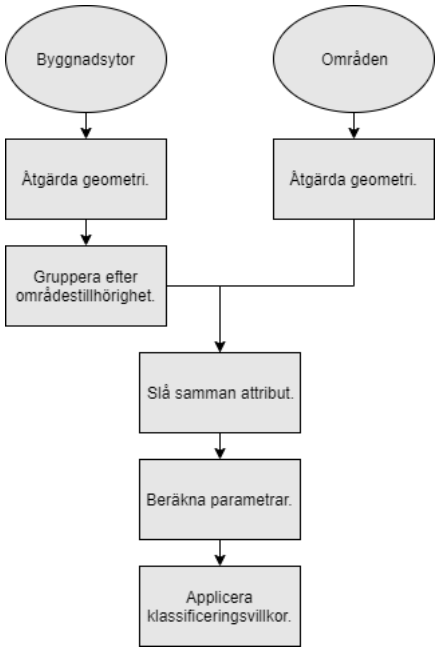
För att beräkna minimum- och maximum-värden för dessa parametrar, parades rätt byggnadsytor ihop med rätt områden i FME Workbench. Separata vektordata innehållande polygoner för byggnadsytor och fastighetsområden importerades, och alla enheters geometri genomgick ett antal tester för att försäkra att de var lämpliga att använda för beräkning. Transformatorer som *Arcstroker*, *Geometryfilter* och *Geometryvalidator* användes först för att säkerställa att alla enheter var i polygonform och reparera de enheter som kunde orsaka eventuella problem för areaberäkning i händelse av att de inte var hela. Till exempel förekom polygoner som inte var sammanhållande eller av varierande geometrityp. Eventuella dubletter sorterades bort med *DuplicateFilter*. Alla byggnadsytor grupperades sedan utefter vilket område de låg inom med hjälp av transformatorn *Clipper*. Arean för alla ytor beräknades med hjälp av *AreaCalculator*. De grupperade byggnadsytorna assignerades sedan ett identifikationsnummer beroende på vilket område de tillhörde genom att slå samman attributen tillhörande både områdena och byggnadsytorna med *FeatureJoiner*.

Sedan kunde alla parametrar och deras minimum-, maximum-värden beräknas med hjälp av *StatisticsCalculator*. All beräknad statistik sammanställdes med *Featurejoiner* och grupperades efter tidigare manuellt utförda klassificeringar med avseende på bebyggelsetyp för att utgöra träningsdata som kunde användas för att utföra klassificeringar. Statistiken exporterades slutligen till ett kalkylblad i Microsoft Excel för att utgöra klassificeringsvillkoren. Dessa kan ses i sin helhet i Bilaga B.

3.3.2 Klassificeringsmetod

För att utföra klassificeringen utgick arbetet från att bebyggelsetypen för varje område var okänt, till skillnad från när villkoren erhölls. Processen liknar de beskrivna under Avsnitt 3.3.1 till en början. De transformatorer som användes finns listade i Bilaga A. Metoden går ut på att återigen gruppera alla byggnadsytor efter vilket område de tillhör, beräkna de ovan beskrivna parametrarna, och sedan applicera villkor baserade på tidigare beräknad statistik för varje bebyggelsetyp.

Ett förenklat flödesschema som representerar processen finns i Figur 13. Polygoner för byggnadsytor och områden importerades, och all geometri sågs över med *Arcstroker*, *Geometryfilter* och *Geometryvalidator*. Dubbletter sorterades bort med *DuplicateFilter* och alla byggnadsytor grupperades efter området de tillhörde med *Clipper*. Parametrarna beräknades sedan med *StatisticsCalculator* och *ExpressionEvaluator*.



Figur 13. *Förenklat flödesschema över metoden för klassificering.*

När alla parametrar beräknats, utvärderades ett antal villkor i *Attributemanager* genom att döpa om ett attribut med hjälp av funktionen *Conditional Value*, eller *Villkorligt värde*. Varje klassificeringsalternativ - alltså varje bebyggelsetyp, fick ett antal villkor som varje enhet var tvungen att uppfylla för att erhålla motsvarande klassificering. Dessa formulerades i som klassificeringsvillkor. Alla minimum- och maximum-värden som används i villkoren kommer från Excel-filen, i vilken statistiken för alla parametrarna listades efter bebyggelsetyp. Konstruktionen av villkoren illustreras i Tabell 9. Klassificeringen lagrades som ett attributvärde, och resulterande enheter exporterades som nytt vektordata.

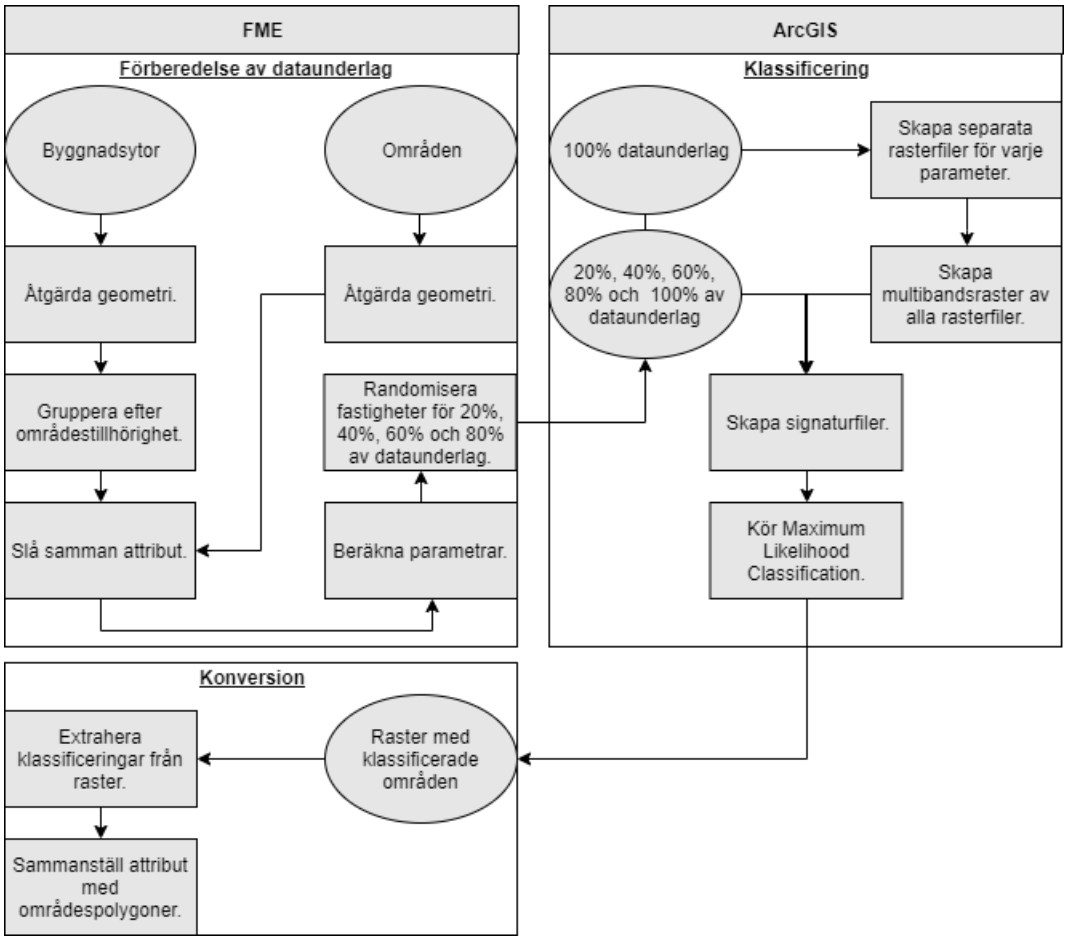
Tabell 9. *If-loop som användes som villkorstest för att ange klassificering som attributvärde för varje område.* i_1–i_n *är de olika parametrarna som testades, och* j_1–j_n *är de olika bebyggelsetyperna som utgör klassificeringsalternativen.*

	Testvillkor	**Attributvärde**
IF	Parameter $i_1 \geq$ minimum för i_1 för klass j_1 AND Parameter $i_1 \leq$ maximum för i_1 för klass j_1 ... AND Parameter $i_n \geq$ minimum för i_n för klass j_1 AND Parameter $i_n \leq$ maximum för i_n för klass j_1	j_1
ELSE IF	...	j_i
ELSE	Parameter $i_1 \geq$ minimum för i_1 för klass j_n AND Parameter $i_1 \leq$ maximum för i_1 för klass j_n ... AND Parameter $i_n \geq$ minimum för i_n för klass j_n AND Parameter $i_n \leq$ maximum för i_n för klass j_n	j_n

3.4 METOD II - KLASSIFICERING BASERAD PÅ STÖRSTA SANNOLIKHET

Den ursprungliga idéen bestod i att utveckla ett poängsystem som poängklassade ett område ju närmre det låg medelvärdet för en viss bebyggelsetyp med avseende på en parameter, för att därefter klassificera området baserat på vilket medelvärde den ligger närmast. I stora drag är det precis vad klassificeringsmetoden baserad på största sannolikhet gör. Detta för att kringgå eventuella problem med överlappning mellan parametrar för olika bebyggelsetyper.

Både FME dekstop och ArcGIS användes, då dataunderlaget först behövde förberedas, samt för att utnyttja ett redan existerande verktyg för klassificering baserad på största sannolikhet i ArcGIS. Metoden byggde på att utföra klassificeringen baserad på samma parametrar baserade på byggnadsytor som under Avsnitt 3.3.1, men att lagra denna information i rasterceller i syfte att kunna använda klassificeringsverktyget **Maximum Likelihood Classification** i ArcGIS. Ett multibandsraster skapades av samtliga raster innehållande parametrar vilka skulle analyseras, och detta användes för att utföra alla klassificeringar. Olika andelar av det ursprungliga dataunderlaget användes i syfte att undersöka hur väl klassificeringarna utfördes med tillgång till varierande dataunderlag. Ett förenklat flödesschema som illustrerar processen finns i Figur 14. Samtliga transformatorer som användes och deras beskrivningar finns listade i Bilaga A.



Figur 14. *Förenklat flödesschema över metoden för klassificering baserad på största sannolikhet.*

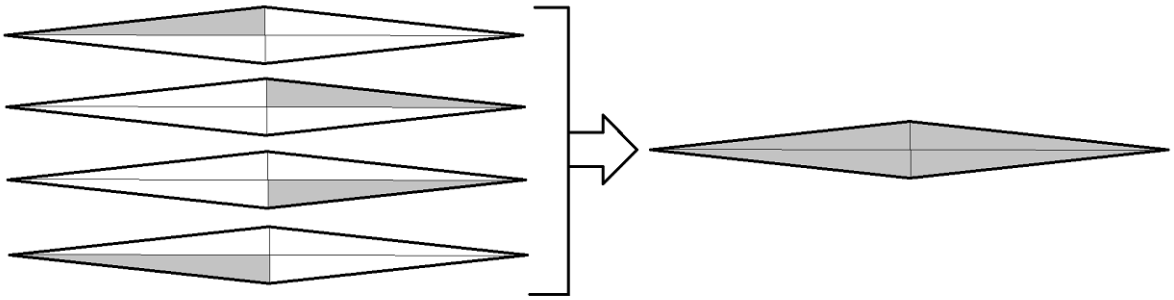
3.4.1 Förberedelse av dataunderlag och parametrar baserade på byggnadsytor

Dataunderlaget förbereddes i FME. Till en början förbereddes dataunderlaget på samma sätt som i övriga metoder. All geometri sågs över med *ArcStroker*, *Geometryfilter* och *GeometryValidator*. Dubbletter togs bort med *Duplicatefilter* och alla byggnadsytor grupperades efter vilket område de låg inom med hjälp av *Clipper*. Därefter beräknades alla parametrar med *StatisticsCalculator* och *ExpressionEvaluator*, och alla attribut slogs samman. Därefter kontrollerades det hur alla parametrar som skulle analyseras förhöll sig till en normalfördelningskruva, då detta är en förutsättning för att klassificering baserad på största sannolikhet ska fungera. Detta gjordes i RStudio, och de parametrar som inte var normalfördelade logaritmerades och kontrollerades återigen. Resultatet var en shapefil innehållande polygoner som motsvarade alla områden, samt alla parametrar som skulle analyseras lagrade som attribut.

För att skapa träningsdata förbereddes det sedan tidigare klassificerade dataunderlaget genom att randomisera fastigheter för 20%, 40%, 60% och 80% av dataunderlaget. Detta gjorde genom att generera jämnt fördelade randomiserade tal för hela dataunderlaget med *RandomNumberGenerator*, och sedan dela in dem i olika grupper med hjälp av *Testfilter*. På så sätt fick underlaget en jämn och randomiserad spridning. Dessa för att simulera att förberedandet av ett underlag genom att slumpmässigt klassificera områden med en så jämn utspridning som möjligt - därmed skapandet av träningsdata för klassificeringsmetoden.

3.4.2 Klassificering

När 20%, 40%, 60% och 80%, samt 100% (den ursprungliga shapefilen med alla polygoner) av dataunderlaget hade förberetts, skapades rasterfiler för varje parameter utifrån vad som utgjorde 100% av dataunderlaget med verktyget *Polygon to Raster*. Därefter skapades ett multibandsraster av samtliga rasterfiler med verktyget *Composite Bands*. På så sätt erhölls ett multibandsraster bestående av rasterfiler där varje cell i varje raster innehöll parameterinformation för det område cellen befann sig i, på samma sätt som illustreras i Figur 15. En lämplig upplösning valdes med cellstorlek 10 för att cellerna inte skulle bli för stora och överlappa fler områden, utan istället falla inom separata områden.



Figur 15. *Hur ett flertal rasterfiler sammanställs till ett multibandsraster. I detta fall sammanställdes multibandsrastret av olika parameterdata.*

Sedan skapades signaturfiler med hjälp av 20%, 40%, 60% och 80% och 100% av dataunderlaget genom att köra verktyget *Create Signatures* på multibandsrastret. På så sätt beräknas kovariansen och medelvärdet för alla parametrar och dessa kan relateras till varandra. Då kunde det även undersökas hur mycket av dataunderlaget som var nödvändigt för att skapa den signaturfil som behövdes för att klassificera alla områden. Efter att signaturfilerna hade skapats så körde verktyget *Maximum Likelihood Classification* på multibandsrastret med de olika signaturfilerna, och resulterade i nya genererade rasterfiler där alla celler hade tilldelats en klassificering i form av en bebyggelsetyp.

3.4.3 Konversion

För att kunna utvärdera hur många områden eller fastigheter som hade klassificerats rätt behövde alla raster med klassificerade celler konverteras tillbaka till de polygoner som motsvarade alla områden, och cellernas klassificeringar behövde överföras till motsvarande polygoner. Detta gjordes enklast i FME Workbench.

För att erhålla rastercellernas klassificeringar i attributform användes en transformator som kallas för *PointonRasterValueExtractor*, vilken extraherar bandvärden ur en rasterfil vid en särskild punkt. Eftersom att den behöver både en rasterfil och punkter som input, behövde punkterna skapas. Detta gjordes med *RasterCellCoercer*, som användes för att skapa en punkt i centrum av varje rastercell. Klassificeringarna extraherades därefter ur rasterfilerna med hjälp av dessa punkter, vilket resulterade i lika många enheter som det fanns tillgängliga punkter med klassificeringarna lagrade som attributvärden. För att erhålla motsvarande och ursprungliga polygoner så sorterades först dubbletter bort med *duplicatefilter*. Sedan användes *Clipper* med punkterna och de ursprungliga polygonerna för att sortera alla punkter gruppvis efter vilket område de tillhörde. I samma transformator så slogs attribut och geometri samman. Därmed erhölls de ursprungliga polygonerna, fast nu med klassificeringar baserade på största sannolikhet lagrade som attributvärden.

3.4.4 Parametrar baserade på servisförekomst

Arbetsprocessen för klassificering baserad på största sannolikhet är ganska omfattande. Åtminstone när flera parametrar används. Därför valdes ett något enklare sätt för att utvärdera några ytterligare parametrar baserade på servisförekomst för metoden. För att utvärdera ett fåtal enklare parametrar så är det ofta tillräckligt att undersöka deras respektive täthetsfunktioner för olika klasser. För parametrar baserade på servisförekomst framställdes därför täthetsfunktioner i RStudio. De parametrar som undersöktes var:

- Antal dagvattenserviser inom ett område
- Antal spillvattenserviser inom ett område
- Kvoten mellan antalet dagvattenserviser inom ett område och dess area
- Kvoten mellan antalet spillvattenserviser inom ett område och dess area.

För att skapa täthetsfunktionerna behövdes först parametrarna beräknas. Detta gjorde genom att att använda transformatorerna *SpatialFilter* i FME Workbench för att ta reda på vilka serviser som låg inom vilka områden. Med *StatisticsCalculator* beräknades antalet spill- och dagvattenserviser inom varje område. Kvoten mellan dessa och områdets area beräknades sedan med *ExpressionEvaluator*. Samtliga enheter slogs sedan samman med *FeatureJoiner*. Med transformatorn *TestFilter* sorterades alla områden utefter vilken bebyggelsetyp de tillhörde, för att sedan skrivas till separata filer. Dessa filer importerades i RStudio där täthetsfunktionerna skapades med kommandot *density* och sammanställdes i grafer.

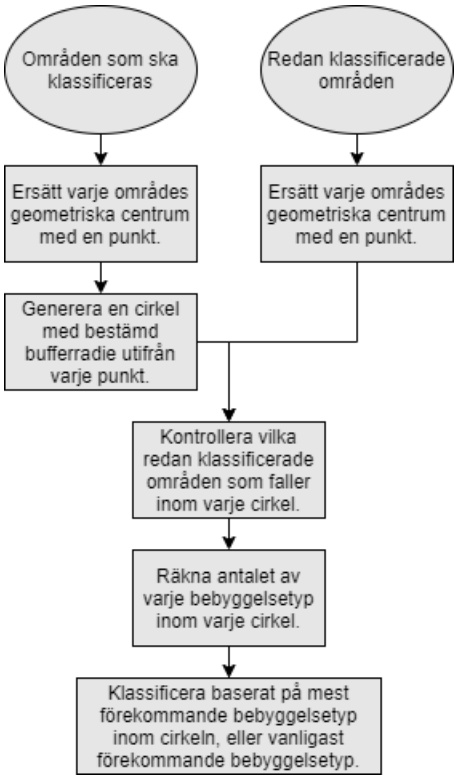
Genom att undersöka om alla parametrars täthetsfunktioner för samtliga bebyggelsetyper överlappar var det möjligt att bestämma huruvida dessa utgjorde lämpliga parameterval eller inte. Om det finns tillräckligt stora skillnader mellan åtminstone en del olika bebyggelsetyper så innebär det att de tillsammans skulle kunna skilja klasserna åt genom klassificeringsmetoden baserad på största sannolikhet.

3.5 METOD III - NÄRMASTE GRANNE-ALGORITMEN OCH INTILLIGGANDE OMRÅDEN

FME Desktop användes för att utveckla en metod baserad på en närmaste granne-algoritmen. Möjligheten att kunna klassificera områden baserad på intilliggande områdens bebyggelsetyp förutsätter att en del av områdena redan har klassificerats. Därför delades metoden upp i två delar. I en del testades algoritmen under olika villkor, med olika parametrar och med varierande andel tillgängligt dataunderlag (sedan tidigare klassificerade områden). I en andra del optimerades villkoren i rimlig utsträckning, men testades fortfarande med varierande andel tillgängligt dataunderlag. På så sätt kunde metoden utvärderas genom att undersöka hur stor andel av dataunderlaget som behöver klassificeras på förhand för att uppnå ett visst resultat, samt hur dessa resultat påverkas av algoritmens villkor och parametrar. Andelen sedan tidigare klassificerade områden varierades mellan 5%, 10%, 20%, 40%, 60% och 80%.

3.5.1 Algoritm utan distansviktning

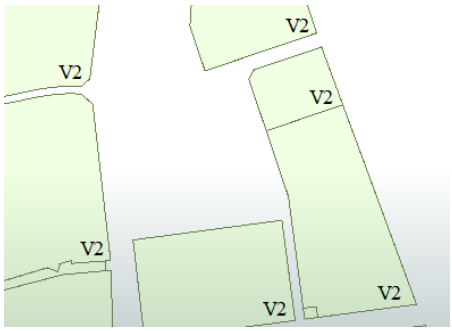
För att utföra klassificeringen baserat på intilliggande områden behövdes det först och främst fastställas vad för typer av områden som fanns i närheten av varje enskilt område. Därefter behövde klassificeringen göras baserat på vilken bebyggelsetyp som var mest förekommande inom en viss närhet av varje område som skulle klassificeras. Flödesschemat i Figur 16 beskriver processen i övergripande detalj.



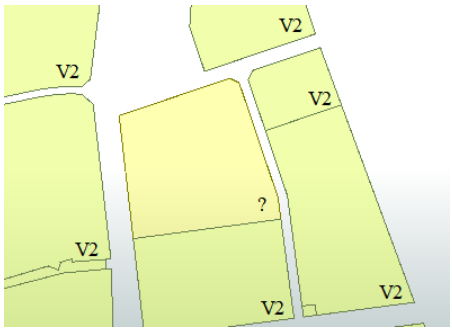
Figur 16. *Förenklat flödesschema över metoden för klassificering baserad på närmaste granne-algoritmen.*

Transformatorn *CenterPointReplacer* användes för att ersätta varje område med en punkt i dess geometriska centrum. *GeographicBufferer* användes sedan för att skapa en buffercirkel med en viss radie kring varje punkt som motsvarade ett område som skulle klassificeras. För att se vilka punkter som låg inom varje cirkel användes sedan *PointOnAreaOverlayer*. Denna transformator undersökte alla punkter inom varje cirkel och lagrade antalet överlappningar som attribut, samt skapade listor som innehöll vilken bebyggelsetyp varje punkt hade. Informationen i dessa listor hämtades med hjälp av *ListExploder*, och antalet punkter av en särskild bebyggelsetyp inom varje cirkel beräknades med *StatisticsCalculator*. Antalen lagrades i en kolumn för varje bebyggelsetyp, och allt grupperades utefter område med hjälp av *Aggregator*. Genom att formulera villkor i *Attributemanager* kunde varje område klassificeras baserat på vilken bebyggelsetyp som var mest förekommande. Om det fanns lika många av olika bebyggelsetyper så klassificerades området baserat på vilken bebyggelsetyp som är vanligast förekommande baserat på underlaget presenterat i Tabell 6 under Avsnitt 3.1.1.

Processen illustreras visuellt i Figur 17, där två punkter vars korresponderande områden är klassificerade med bebyggelsetypen V2 föll inom cirkeln för området som ska klassificeras. Hade radien utökats med bara några meter, hade det varit fyra V2-punkter. Inga andra bebyggelsetyper finns inom cirkeln och därför klassificeras även detta område som ett V2-område.

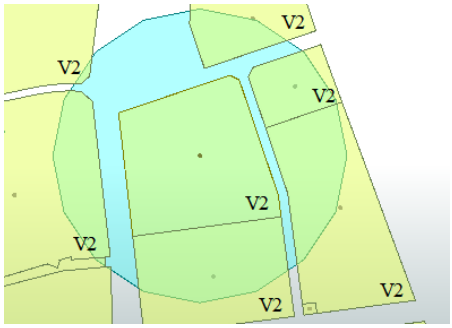


(a). *Redan klassificerade områden.*



(b). *Ett icke klassificerat område bland redan klassificerade områden.*

(c). *Punkter för alla områden.*



(d). *Buffercirkel utifrån punkten för området som ska klassificeras.*

Figur 17. *De olika stegen i en klassificeringsmetod baserad på vilka områden som finns inom en viss radie från dess centrum.*

För att undersöka hur cirklarnas radie påverkade resultatet testades metoden med radierna 50, 100, 200, 400, 600 och 800 meter. Arbetet utgick ifrån att användandet av områden som ligger närmre är troligare att bidra till en mer korrekt klassificering än områden som ligger längre bort. Däremot kan en del områden exkluderas om radien är för liten, då det inte nödvändigtvis faller redan klassificerade områden inom varje cirkel. Därför undersöktes metodens prestation i sin helhet, samt för varje bebyggelsetyp för varje test. Även andelen områden som klassificerats i varje test jämfördes.

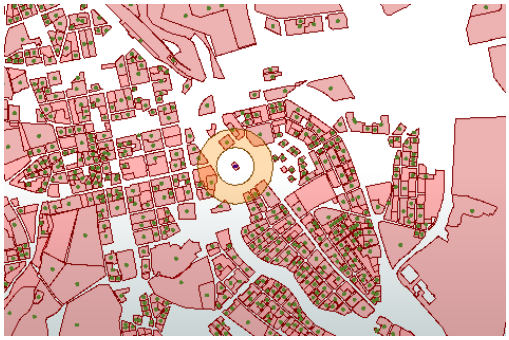
3.5.2 Algoritm med distansviktning

För att garantera att alla områden klassificerades utvecklades metoden till att utöka buffercirklarnas radie tills alla områden omfattades. Detta gjordes genom att utgå från samma metod, men även genom att skapa fler buffercirklar. Betydelsen för klassificering blev mindre ju större cirkeln var, då områden som låg längre bort skulle få mindre betydelse för klassificeringen än områden som låg närmre. De olika storleksintervallen som användes för cirklarna var 50, 100, 200, 400, 600 och 800 meter. Först skapades buffercirklar med dessa radier. Föregående mindre buffercirkelar klipptes ur de större cirklarna med hjälp av transformatorn *Clipper* för att skapa cirkelsegment och på så sätt inte inkludera samma områden flera gånger. Detta illustreras i Figur 18, där området inte kan klassificeras förrän

i steget som visas i Figur 18c då den yttre radien har utökats till 200 meter.

(a). *Buffercirkel med radie 50 meter.*

(b). *Buffercirkelsegment med inre radie 50 meter och yttre radie 100 meter.*



(c). *Buffercirkelsegment med inre radie 100 meter och yttre radie 200 meter.*

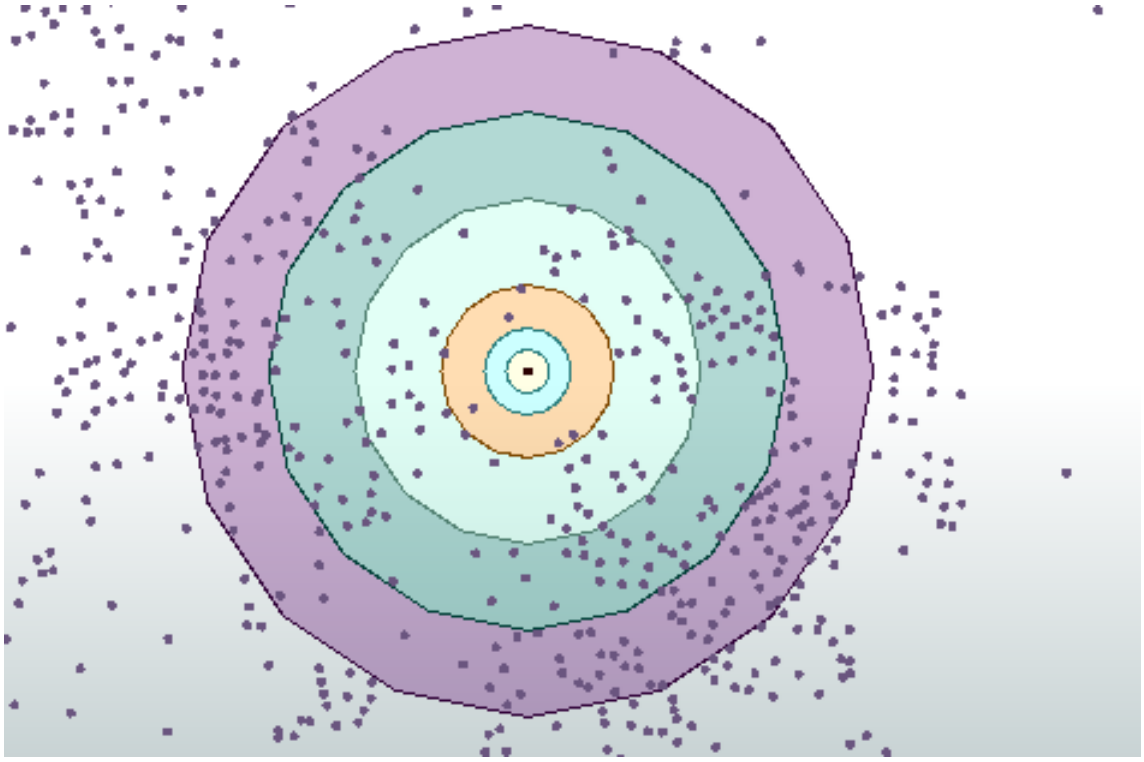
(d). *Buffercirkelsegment med inre radie 200 meter och yttre radie 400 meter.*

(e). *Buffercirkelsegment med inre radie 400 meter och yttre radie 600 meter.*

(f). *Buffercirkelsegment med inre radie 600 meter och yttre radie 800 meter.*

Figur 18. *Ett område som ska klassificeras bland ett flertal redan klassificerade områden och de olika buffercirklar och segment som användes.*

Detta framgår ännu tydligare om alla områdesgränser tas bort och bilden blir lite större, vilket visas i Figur 19. Då syns det att sex stycken punkter hamnar inom den tredje buffercirkeln från mitten. På så sätt utökas cirkelns storlek tills alla områden har blivit klassificerade. Varje område som ska klassificeras genomgår samma process i algoritmen.



Figur 19. *Ett område som ska klassificeras bland ett flertal redan klassificerade områden och de olika buffercirklar och segment som användes, utan områdesgränser.*

För att försäkra att områden närmre området som skulle klassificeras viktades tyngre, infördes en faktor som multiplicerades med antalen bebyggelsetyper inom varje buffercirkel och segment. På så sätt garanterades en så hög överenstämmelsegrad som möjligt. Multiplikationsfaktorn för varje bufferelement som användes finns i Tabell 10. På så sätt klassificeras varje område utefter det högsta antalet av en viss bebyggelsetyp inom ev viss närhet av området samtidigt som metoden garanterar att alla områden klassificeras.

Tabell 10. *Multiplikationsfaktorer för antalet bebyggelsetyper inom olika buffercirklar och segment.*

Inre radie [m]	Yttre radie [m]	Faktor
0	50	1
50	100	0,1
100	200	0,01
200	400	0,001
400	600	0,0001
600	800	0,00001

3.6 METOD IV - KLASSIFICERING AV ANSLUTNINGSFÖRHÅLLANDEN

För att skilja på områden med olika anslutningsförhållanden valdes en enklare metod med anledning av att dataunderlaget inte tillät en utvärdering av metoden. Dataunderlaget från Linköping kunde ej uvärderas eftersom att inga av områden har anslutningsförhållanden som attribut, och det fanns ej dag- eller spillvattenledningar att tillgå i Västervik. Eftersom att det fanns flest områden att arbeta med i Linköping valdes detta dataunderlag för att utveckla metoden. Metoden validerades endast genom visuell inspektion. Eftersom att en del ledningar och serviser inte alltid är helt korrekt representerade delades metoden upp i två delar. En där metoden testades utan att manipulera dataunderlaget, och en del där metoden testades efter att ha genomfört en del förändringar.

Metoden ämnade att kunna skilja på områden med anslutningsförhållanden enligt vad som förklaras under Avsnitt 2.1.3, men begränsades till kategorierna som omfattas i Figur 2 - alltså *S*, *D* och *K*. Resterande valdes att kategoriseras som okända, då det troligtvis endast handlar om ett fåtal områden, vilka istället kommer att kräva en visuell bedömning. För att skilja på områdena användes FME Workbench och följande villkor för följande klasser:

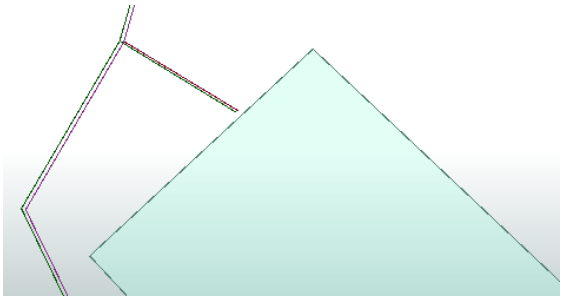
S: Om det finns både spill- och dagvattenserviser inom området, samt spill- och dagvattenledningar i intilliggande gata så är det ett område med separerade ledningsnät.
D: Om det inte finns någon dagvattenservis inom området, samt även en dagvattenledning i intilliggande gata så är det ett område med delvis kombinerade ledningssystem.
K: Om det inte finns någon dagvattenledning i intilliggande gata så är det ett område med kombinerade ledningssystem.

Områden som kategoriserades som okända delades upp i ytterligare fyra underkategorier: *O*, *O1*, *O2* och *O3*. Detta i syfte att förenkla eventuellt efterarbete. Dessa motsvarar:

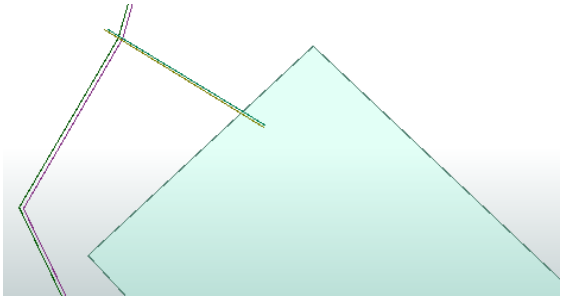
O: Om det inte finns några serviser inom området, samt inga ledningar i intilliggande gata.
O1: Om det inte finns några serviser inom området, men spillvattenledningar i intilliggande gata.
O2: Om det inte finns några serviser inom området, men dagvattenledningar i intilliggande gata.
O3: Om det inte finns några serviser inom området, men både spill- och dagvattenledningar i intilliggande gata.

Metoden delades upp i två delar - med och utan datamanipulation. I delen med datamanipulation så utvecklades metoden ytterligare genom att ge utrymme för förändringar av dataunderlaget. För att kringgå problemet med att en del serviser inte är ritade hela vägen fram till vissa fastigheter så förlängdes samtliga serviser fem meter. Att förlänga dem mycket längre än så skulle kunna ge upphov till problem eftersom att en del serviser i enskilda fall då kunde överlappa med fel fastigheter, medan en sträcka mindre än så inte tycktes ge upphov till större skillnader. Serviser förlängdes med hjälp av transformatorn *LineExtender*. På så sätt skulle eventuella felaktiga klassificeringar med avseende på anslutningsförhållanden kunna undvikas.

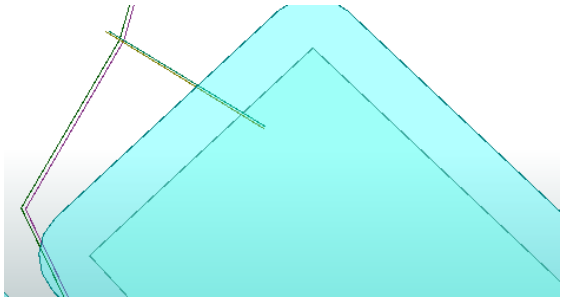
För att avgöra vilka ledningar och serviser som föll inom vardera område och dess intilliggande gator användes transformatorn *SpatialRelator*. Först användes områdena som filter och de förlängda serviserna som kandidater. Sedan skapades bufferzoner på 15 meter kring varje område för att dessa skulle kunna tillgodoräkna sig ledningar som låg i intilliggande gator. Dessa bufferzoner användes som filter tillsammans med spill- och dagvattenledningarna som kandidater för att beräkna antalet ledningar i varje intilliggande gata. Detta illustreras i Figur 20.



(a). *Ursprungligt dataunderlag - (0 spillvattenserviser, 0 dagvattenserviser, 0 spillvattenledningar, 0 dagvattenledningar)*



(b). *Dataunderlag efter förlängningar - (1 spillvattenservis, 1 dagvattenservis, 0 spillvattenledningar, 0 dagvattenledningar).*



(c). *Dataunderlag med förlängningar och bufferzon kring ett område - (1 spillvattenservis, 1 dagvattenservis, 1 spillvattenledning, 1 dagvattenledning).*

Figur 20. *Ett exempel på hur metoden tillämpas på ett område med separerade anslutningsförhållanden.*

Antalen enheter inom varje område beräknades därmed separat för ledningar och serviser. *AttributeCreator användes* för att skapa attribut innehållande nollor för de områden som saknade ledningar eller serviser. Efter att samtliga antal hade beräknats slogs dessa samman med transformatorn *Aggregator*. Vid detta steget hade alltså varje enhet (område) antalet dagvattenledningar i intilliggande gata, antalet spillvattenledningar i intilliggande gata, antalet dagvattenserviser inom området och antalet spillvattenserviser inom området lagrat som attribut. Dessa enheter testades sedan med tidigare beskrivna villkor med hjälp av *TestFilter*.

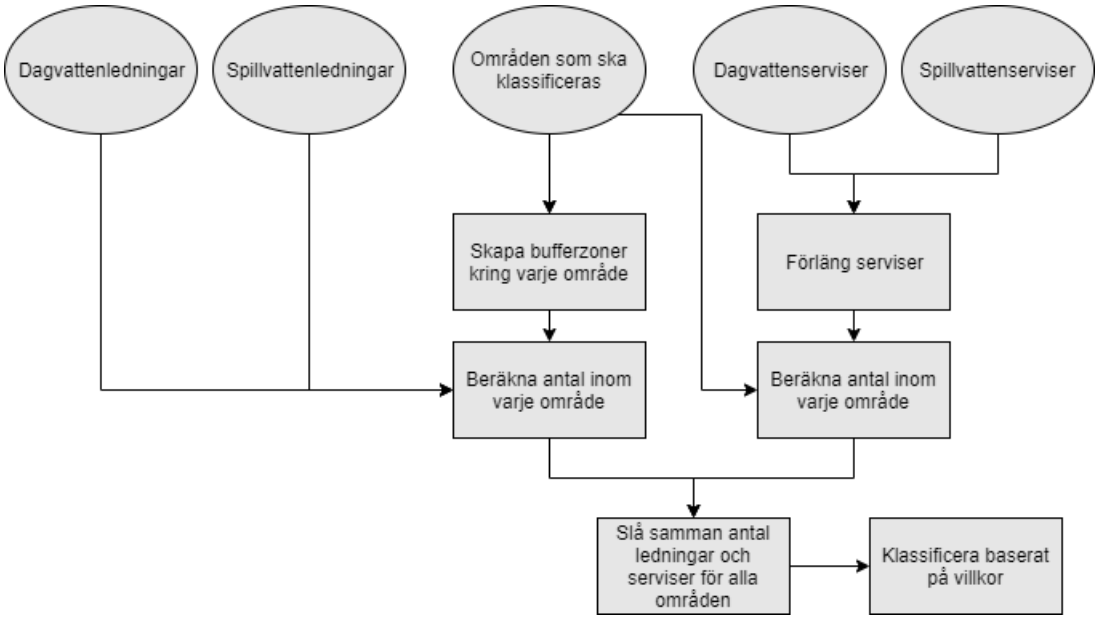
Villkoren formulerades som en IF-loop och går att se i Tabell 11. Först och främst klassificerades områden utan några serviser som okända, för att sedan kunna klassificerades

på ett annorlunda sätt eller genom visuell inspektion. Områden med kombinerade system klassificerades sedan baserat på förutsättningen att inga dagvattenserviser förekom inom området, samt att det inte heller fanns några dagvattenledningar i gatan. För delvis kombinerade områden förutsattes det att det fanns åtminstone en dagvattenledning i intilliggande gata, men inga dagvattenserviser inom området. Resterande områden antogs ha separerade anslutningsförhållanden eftersom att kvarvarande områden hade någon kombination av lednings- och servisförekomst där samtliga antal var större än noll.

Tabell 11. *Villkor för klassificering med avseende på anslutningsförhållanden formulerade för FME Workbench i form av en IF-loop.*

	Testvillkor	**Attributvärde**
IF	Antalet dagvattenledningar $= 0$ AND Antalet dagvattenserviser $= 0$ AND Antalet spillvattenledningar $= 0$ AND Antalet spillvattenserviser $= 0$	O
ELSE IF	Antalet dagvattenledningar $= 0$ AND Antalet dagvattenserviser $= 0$ AND Antalet spillvattenledningar > 0 AND Antalet spillvattenserviser $= 0$	O1
ELSE IF	Antalet dagvattenledningar > 0 AND Antalet dagvattenserviser $= 0$ AND Antalet spillvattenledningar $= 0$ AND Antalet spillvattenserviser $= 0$	O2
ELSE IF	Antalet dagvattenledningar > 0 AND Antalet dagvattenserviser $= 0$ AND Antalet spillvattenledningar > 0 AND Antalet spillvattenserviser $= 0$	O3
ELSE IF	Antalet dagvattenledningar $= 0$ AND Antalet dagvattenserviser ≥ 0 AND Antalet spillvattenledningar ≥ 0 AND Antalet spillvattenserviser > 0	K
ELSE IF	Antalet dagvattenledningar > 0 AND Antalet dagvattenserviser $= 0$ AND Antalet spillvattenledningar ≥ 0 AND Antalet spillvattenserviser ≥ 0	D
ELSE		S

På så sätt erhöll varje område en klassificering med avseende på anslutningsförhållanden. Dessa sorterades sedan med *Attributemanager* och erhöll åter sin geometri med *FeatureMerger* då områdenas geometri försummades vid beräkningar. Ett överskåligt flödesschema av processen går att se i Figur 21.



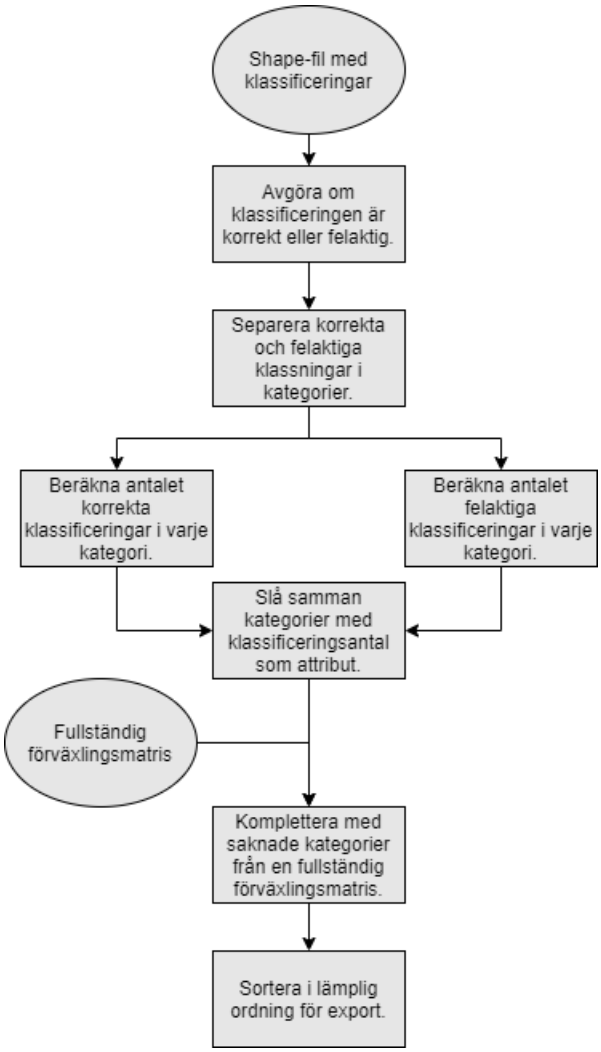
Figur 21. *Flödesschema över processen för att klassificera ett område med avseende på anslutningsförhållanden med dataunderlagsmanipulation.*

3.7 UTVÄRDERING OCH JÄMFÖRELSE AV METODER

3.7.1 Export av prestationsdata

För att utvärdera resultaten av olika klassificeringsmetoder med avseende på bebyggelsetyp så samlades all information fördelaktigt i en och samma shape-fil. Målet var att samla alla klassificeringar utförda av en modell i en förväxlingsmatris så att dessa sedan kunde jämföras med tidigare manuellt övervakade klassificeringar. Att göra detta innebar ett flertal steg, vilka gjordes i mjukvaran FME Desktops Workbench. Samtliga transformatorer som användes för att exportera den data som ska användas till att undersöka klassificeringsmodellernas prestation är listade med överskådliga beskrivningar i Bilaga A.

Först och främst behövdes en fil, eller filer, som innehåller de klassificeringar som har gjorts av modellen i attributform. Dessa behövde ordnas och föras in på rätt plats i en förväxlingsmatris. På så sätt kunde även mindre modellförändringar snabbare utvärderas eftersom att det är mycket tidskrävande att manuellt föra in antalet gissningar inom olika kategorier i en större matris. En konceptuell modell av hur detta utfördes finns i form av ett flödesschema i Figur 22.



Figur 22. *Flödesschema över exportmetoden för utvärdering av klassificeringar.*

För att avgöra om varje klassificering var korrekt eller felaktig användes först transformatorn *Tester*. Testvillkoret som angavs var huruvida innehållet i attributet som innehöll modellens klassificering var lika med innehållet i attributet som innehöll den på förhand bestämda klassen. På så sätt grupperades alla korrekta gissningar i en grupp, och alla felaktiga gissningar i en annan. Därefter så behövdes alla enheter grupperas ytterligare. Detta eftersom att utöver huruvida en klassificering var rätt eller fel, är det av intresse att veta precis vad modellen klassificerade korrekt och inkorrekt. Till exempel om en modell endast klassificerade en klass korrekt, eller hur många gånger modellen klassificerade en särskild klass felaktigt.

För den grupp där klassificeringen var korrekt angavs villkoren på följande sätt. Alla korrekta klassificeringar grupperades utefter klass. Alltså hamnade alla korrekta enheter för en klass i en grupp. De fullständiga testvillkoren för gruppen med korrekta klassificeringar finns att se i Bilaga E - Export av prestationsdata. För de felaktiga klassificeringarna blev testvillkoren betydligt fler eftersom att det krävdes en grupp för varje enskild kombi-

nation av modellklassificering och egentlig klassificering. En grupp består till exempel av alla enheter modellen har klassificerat som A, men egentligen var av klass B. Nästa grupp består av alla enheter modellen har klassificerat som A, men egentligen var av klass C. Så fortsätter det för alla kombinationer av felaktiga klassificeringar.

I nästa steg användes transformatorn *StatisticsCalculator* för att beräkna antalet gånger varje klassificeringskombination förekom, och därefter gruppera dem. Istället för en lista med alla enheter, erhölls därmed en lista med alla klassificeringskombinationer och deras förekomst. Dessa kunde sedan kombineras med hjälp av transformatorn *AttributeManager*. Därmed erhölls en lista på korrekta och felaktiga gissningar. För att föra in den automatiskt i en förväxlingsmatris behövde dock listan kompletteras med alla de eventuella klassificeringskombinationer som ej uppkommit, då en modell nödvändigtvis inte utför alla kombinationer av korrekta och felaktiga kombinationer. Därför förbereddes ett ark i Excel med alla tänkbara kombinationer och dessa angavs antalet noll som attribut för att motsvara icke förekomst. Dessa importerades sedan vid detta steg i FME och kombinerades med hjälp av transformatorn *Aggregator*, efter att ha manipulerats än en gång med *AttributeManager*. Slutligen sorterades alla enheter, vilka nu motsvarade alla tänkbara kombinationer av klassificeringar, i bokstavsordning. Sedan exporterades de till Excel direkt in i en mall för en förväxlingsmatris baserad på tillgängliga klasser.

3.7.2 Utvärderingsmall för metoder i form av förväxlingsmatris

När alla enheter innehållande kombinationer av klassificeringar hade exporterats från FME till Excel i form av en lista, så kunde dessa enkelt kopplas till en förväxlingsmatris. Då enskilda kombinationers förekomst alltid skrevs ut i samma celler, kopplades dessa celler till rätt plats i matrisen. Med matrisen ifylld med alla kombinationers förekomst, kunde de statistiska måtten beskrivna under Avsnitt 2.4.1 beräknas med formler kopplade till cellerna som utgör förväxlingsmatrisen. I Figur 23 visas en mall för förväxlingsmatrisen som användes.

		SANN														
		C	D	F	FA	FL	FO	GV	I	KO	OB	R	V1	V2	Ö	
UPPSKATTAD	C	150	1	1	2	1	1	2	1	1	1	1	1	1	1	165
	D	1	432	1	1	1	1	1	1	1	1	1	1	1	1	445
	F	3	1	376	1	1	1	1	1	1	1	1	1	1	1	391
	FA	3	1	1	124	1	1	1	1	1	1	1	1	1	1	139
	FL	3	1	1	1	453	1	1	1	1	1	1	1	1	1	468
	FO	1	1	1	1	2	123	1	1	1	1	1	1	1	1	137
	GV	1	1	1	1	1	4	93	1	1	1	1	1	1	1	109
	I	1	1	1	1	1	3	1	32	1	1	1	1	1	1	47
	KO	1	2	1	1	1	2	1	1	546	1	1	1	1	1	561
	OB	1	2	1	2	1	1	1	1	1	122	1	1	1	1	137
	R	1	1	3	2	1	1	1	1	1	1	376	1	1	1	392
	V1	1	1	4	1	1	1	1	1	1	1	1	123	1	1	139
	V2	1	1	2	1	1	1	1	1	1	1	1	1	444	1	458
	Ö	1	1	2	1	1	1	1	1	1	1	1	1	1	321	335
		169	447	396	140	467	142	107	45	559	135	389	136	457	334	

Noggrannhet	0,95
Avvikelse	0,05
Nollratio	0,86
Cohens κ	0,63

F1		
	Medel	0,91
	St.Avv.	0,07
	Varians	0,01
	n	14

	C	D	F	FA	FL	FO	GV	I	KO	OB	R	V1	V2	Ö
Precision	0,91	0,97	0,96	0,89	0,97	0,90	0,85	0,68	0,97	0,89	0,96	0,88	0,97	0,96
Känslighet	0,89	0,97	0,95	0,89	0,97	0,87	0,87	0,71	0,98	0,90	0,97	0,90	0,97	0,96
F1	0,90	0,97	0,96	0,89	0,97	0,88	0,86	0,70	0,98	0,90	0,96	0,89	0,97	0,96

Figur 23. *Mall i Excel för förväxlingsmatris med godtyckliga exempelvärden ifyllda.*

3.7.3 Portabilitet

För att undersöka huruvida det är möjligt att utveckla fungerande alternativa metoder för områdesklassificering undersöktes det även om den bästa metoden gick att implementera på annat dataunderlag. Eftersom att dataunderlaget från Linköping bestod av flest datapunkter, användes främst detta för att utveckla och testa samtliga metoder. För att testa portabiliteten användes dataunderlaget från Västervik.

Den bästa metoden definierades som den metod vilken åstadkom det högsta värdet för Cohens κ. Denna metod testades med dataunderlaget från Västervik. För att kunna kvantifiera portabiliteten hade det egentligen behövts dataunderlag från ett flertal städer som omfattar fler områden med varierande spridning för att utvärdera metoderna. Dock testades den bästa metoden på både Linköping och Västervik i syfte att utgöra ett diskussionsunderlag för hur väl metoden kan tänkas prestera på olika platser och vad en variation i prestation kan bero på.

3.7.4 Tidsåtgång

För att uppskatta tidsåtgången för de olika metoderna vägdes upplevd arbetsbörda mot hur komplicerad metoden var att utveckla utifrån tidigare erfarenheter. Tidsåtgången utvärderades därmed utifrån ett diskussionsunderlag.

4 RESULTAT

4.1 METOD I - REKTANGULÄRKLASSIFICERING OCH BYGGNADSYTOR

Ett överskådligt resultat från rekangulärklassificeringsmetoden med byggnadsytor går att se i Tabell 12. Noggrannheten är mycket låg med denna metoden, och det är ytterst få områden som klassificeras korrekt. Därmed är även avvikelsen mycket hög, och värdet för Cohens κ är långt under noll.

Tabell 12. *Värden för noggrannhet, avvikelse, nollratio och Cohens κ för rektangulärklassificering av bebyggelsetyper.*

Noggrannhet	0,02
Avvikelse	0,98
Nollratio	0,51
Cohens κ	-1,01

Det går att närmre undersöka skillnaden mellan hur olika bebyggelsetyper har klassificerats i Tabell 13. Bebyggelsetypen *F (flerfamiljshus)* har högst F1-värde, men det är fortfarande lågt, och därmed är prestationen inte särskilt bra. För områden av bebyggelsetypen *C (Tomtmark)* så är känsligheten hög, precisionen låg. Detta innebär att dessa områden har klassificerats rätt, men troligtvis bara på grund av att klassificeringen har valts många gånger. Motsatt gäller för områden av typen *I (Industri)*. Klassificeringen har endast gjorts ett fåtal gånger, men den har varit korrekt en stor andel av gångerna eftersom att precisionen är hög.

Tabell 13. *Precision, känslighet och F1-värden för alla klassificeringar, med 20%, 40%, 60%, 80% och 100% av tidigare klassificerade områden använda för att skapa signaturfiler.*

	C	F	FARM	GV	I	R	V1	V2
Precision	0,00	0,30	0,00	0,25	0,83	0,07	0,07	1,00
Känslighet	0,94	0,26	0,00	0,13	0,01	0,00	0,00	0,00
F1	0,01	0,28	0,00	0,17	0,02	0,00	0,00	0,00

4.2 METOD II - KLASSIFICERING BASERAD PÅ STÖRSTA SANNOLIKHET

4.2.1 Parameterval baserade på byggnadsytor

I Tabell 14 presenteras resultatet av Shapiro-Wilks test utförda på de parametrar baserade på byggnadsytor som avseddes användas i metoden. Samtliga parametrar är inte särskilt bra anpassade till normalfördelningar, då de har mycket låga W-värden. Även samtliga p-värden är mindre än 0,05, vilket innebär att nollhypotsen som säger att de är normalfördelade kan förkastas.

Tabell 14. *Resultat av Shapiro-wilks test utförda på alla redan klassificerade områden i Linköping.*

Parameter	**W-värde**	**p-värde**
Antal fastigheter	0,33	$2,2 * 10^{-16}$
Områdets area	0,33	$2,2 * 10^{-16}$
Minsta byggnadsyta	0,15	$2,2 * 10^{-16}$
Största byggnadsyta	0,20	$2,2 * 10^{-16}$
Summa byggnadsyta	0,10	$2,2 * 10^{-16}$
Medel byggnadsyta	0,23	$2,2 * 10^{-16}$

Eftersom att nollhypotsen för samtliga parametrar kunde förkastas så log-transformerades även samtliga parametrar. Resultatet av Shapiro-Wilks test på de logaritmerade värdena kan ses i Tabell 15. Eftersom att ingen paramater är normalfördelad så kvarstår samma p-värde, medan de nu är betydligt bättre anpassade till en normalfördelningskurva. Detta eftersom att W-värdet ligger betyldigt närmre värdet 1 i jämförelse med resultaten i Tabell 14.

Tabell 15. *Resultat av Shapiro-Wilks test utförda på alla redan klassificerade områden i Linköping med logaritmerade värden.*

Parameter	**W-värde**	**p-värde**
Antal fastigheter	0,95	$2,2 * 10^{-16}$
Områdets area	0,93	$2,2 * 10^{-16}$
Minsta byggnadsyta	0,98	$2,2 * 10^{-16}$
Största byggnadsyta	0,81	$2,2 * 10^{-16}$
Summa byggnadsyta	0,93	$2,2 * 10^{-16}$
Medel byggnadsyta	0,85	$2,2 * 10^{-16}$

Värden från förväxlingsmatriser för noggrannhet, avvikelse, nollratio och Cohens κ finns att se i Tabell 16.

Tabell 16. *Värden för noggrannhet, avvikelse, nollratio och Cohens κ för klassificeringsmetoden, med 20%, 40%, 60%, 80% och 100% av tidigare klassificerade områden använda för att skapa signaturfiler.*

	20%	**40%**	**60%**	**80%**	**100%**
Noggrannhet	0,34	0,32	0,29	0,35	0,34
Avvikelse	0,66	0,68	0,71	0,65	0,66
Nollratio	0,48	0,48	0,48	0,48	0,48
Cohens κ	-0,27	-0,32	-0,38	-0,26	-0,28

Noggrannheten och avvikelsen för metoden förefaller vara ganska jämn. Därmed är även värdet för Cohens κ ungefär vid samma nivå oavsett om dataunderlaget består av 20%, 40%, 60%, 80% eller 100% av de tillgängliga redan utförda klassifikationerna. Alla värden för Cohens κ är dock mindre än noll, vilket innebär att metoden inte presterar bättre än om den vanligaste klassificeringen hade valt varje gång. Nollration är konstant vid 0,48 eftersom att den bestäms av kvoten mellan den vanligaste klassificeringen och det totala antalet klassificeringar, vilka inte förändras.

I Tabell 17 finns alla värden för precision, känslighet och F1 för de individuella klassificeringarna. Även här framgår det att metoden presterar på en förhållandevis jämn nivå oavsett hur stort dataunderlaget är.

Tabell 17. *Precision, känslighet och F1-värden för alla klassificeringar, med 20%, 40%, 60%, 80% och 100% av tidigare klassificerade områden använda för att skapa signaturfiler. Värden över 0,50 är markerade med en asterisk (*).*

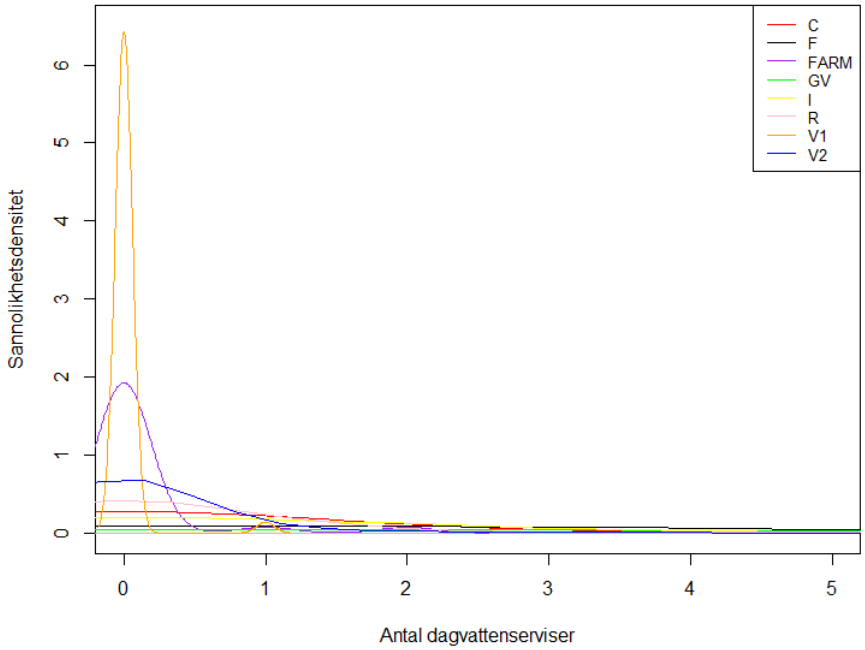
Data ↓	Bebyggelsetyp →	C	F	FA	GV	I	R	V1	V2
20%	Precision	0,02	0,09	0,03	0,01	0,12	0,46	0,02	0,64*
	Känslighet	0,18	0,11	0,11	0,20	0,68*	0,04	0,17	0,55*
	F1	0,03	0,10	0,05	0,02	0,20	0,08	0,03	0,59*
40%	Precision	0,02	0,23	0,00	0,04	0,11	0,52*	0,02	0,62*
	Känslighet	0,06	0,16	0,11	0,13	0,85*	0,03	0,13	0,49
	F1	0,03	0,19	0,01	0,06	0,20	0,05	0,03	0,55*
60%	Precision	0,01	0,13	0,01	0,01	0,15	0,63*	0,01	0,62*
	Känslighet	0,09	0,39	0,14	0,20	0,49	0,02	0,13	0,45
	F1	0,01	0,19	0,01	0,01	0,22	0,05	0,02	0,52*
80%	Precision	0,01	0,17	0,01	0,01	0,11	0,64*	0,02	0,62*
	Känslighet	0,09	0,20	0,04	0,27	0,66*	0,03	0,06	0,57*
	F1	0,01	0,18	0,01	0,02	0,19	0,05	0,03	0,59*
100%	Precision	0,00	0,20	0,00	0,01	0,11	0,64*	0,02	0,63*
	Känslighet	0,06	0,21	0,04	0,33	0,67*	0,02	0,15	0,55*
	F1	0,01	0,21	0,01	0,02	0,19	0,04	0,04	0,59*

Metoden verkar vara bäst på att klassificera bebyggelse av typen *V2*, med precision som rör sig kring 0,60 och känslighet runt 0,50. För bebyggelsetypen *I* är känsligheten högre, medan precisionen är lägre. Detta innebär att metoden är bra på att klassificera rätt när den behöver klassificera bebyggelse av typen *I*, medan den inte har gjort särskilt många korrekta klassificeringar då den har klassificerat områden som *I*. Det motsatta gäller för bebyggelsetypen *R*. Bebyggelsetypen *F* verkar uppnå en känslighet kring 0,2–0,3 först när dataunderlaget överstiger 60%. Fördelningen för de olika klasserna tillhörande datan i Tabell 16 och 17 går att se i färgkodade kartor i Bilaga C.

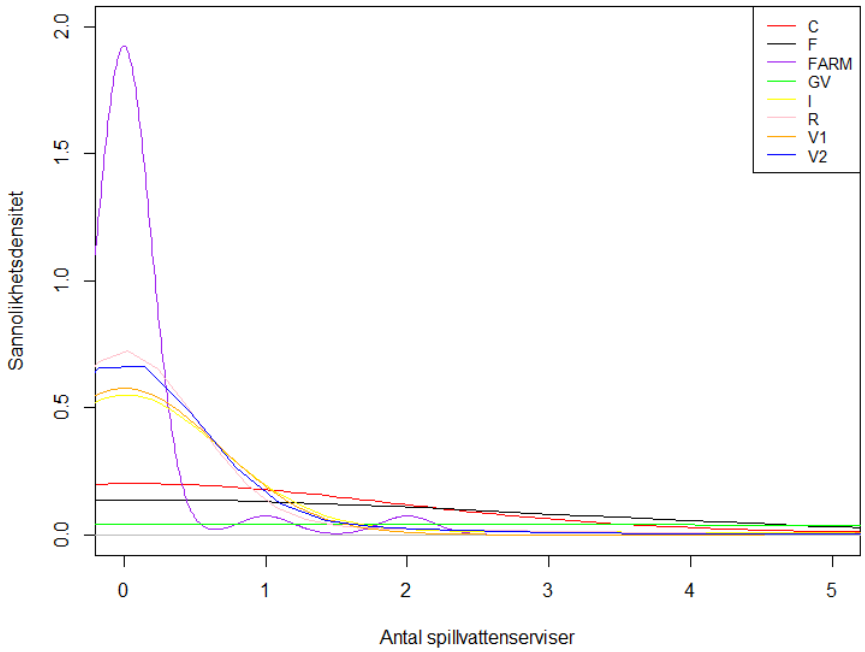
4.2.2 Parameterval baserade på servisförekomst

Sannolikhetsdensiteterna för antalet dag- och spillvattenserviser inom ett område går att se i Figur 24. Det framgår i Figur 24a att sannolikheten för att det inte finns några dagvattenserviser i främst *V1*-områden är större än för någon annan bebyggelsetyp. På samma sätt så syns det i Figur 24b att det sällan finns några spillvattenserviser inom områden av bebyggelsetypen *FARM*. I övrigt så överlappar alla klasser varandra, och är svåra att särskilja. Därmed är troligtvis varken antalet dagvattenserviser eller antalet spillvattenserviser inom ett område lämpligt som parameterval för klassificeringsmetoden baserad på största sannolikhet.

Sannolikhetsdensiteterna mot kvoterna mellan antalet dag- och spillvattenserviser och områdets area syns i Figur 25. Skillnaden mellan bebyggelsetyper för kvoterna är så små att de är svåra att urskilja. Mellan dessa finns det inte heller någon spridning, vilket gör kvoterna mellan antalet dag- eller spillvattenserviser inom ett område och dess area till olämpliga parameterval. Av de totalt fyra parametrarna som undersöktes så förefaller det som om densitetsfunktionerna för samtliga bebyggelsetyper överlappar.

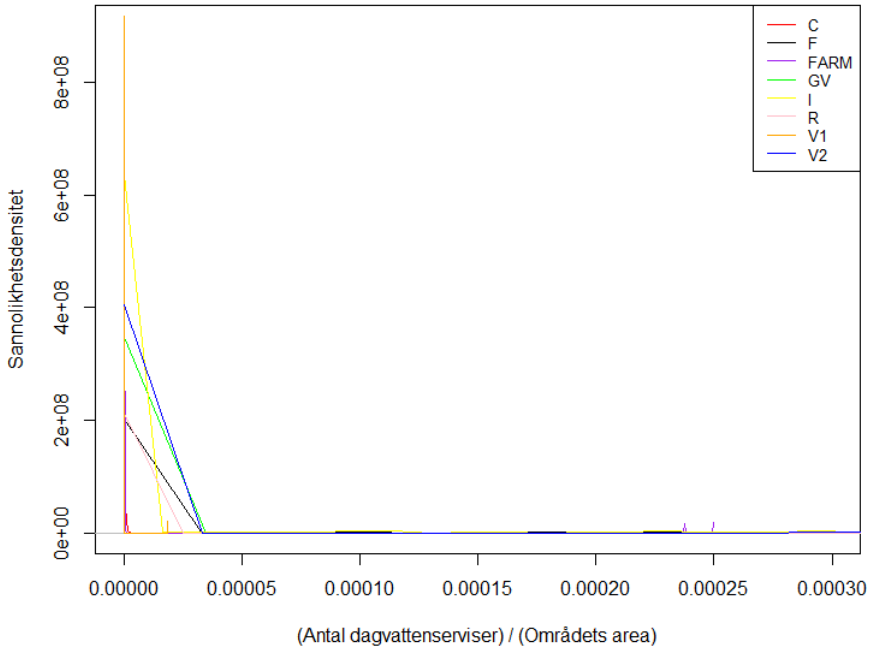


(a). *Sannolikhetsdensitetet med avseende på antalet dagvattenserviser inom ett område av en viss bebyggelsetyp.*

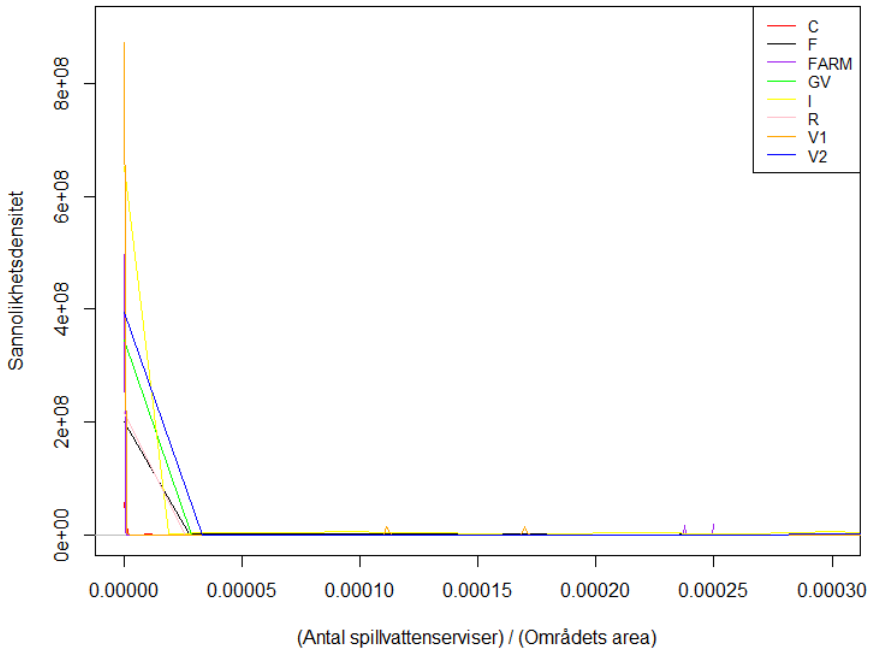


(b). *Sannolikhetsdensitetet med avseende på antalet spillvattenserviser inom ett område av en viss bebyggelsetyp.*

Figur 24. *Sannolikhetsdensitetet med avseende på antalet dag- eller spillvattenserviser inom ett område av en viss bebyggelsetyp.*



(a). *Sannolikhetsdensitetet med avseende på kvoten av antalet dagvattenserviser inom ett område av en viss bebyggelsetyp och dess area.*



(b). *Sannolikhetsdensitetet med avseende på kvoten av antalet dagvattenserviser inom ett område av en viss bebyggelsetyp och dess area.*

Figur 25. *Sannolikhetsdensitetet med avseende på kvoten av antalet dag- eller spillvattenserviser inom ett område av en viss bebyggelsetyp och dess area.*

4.3 METOD III - NÄRMASTE GRANNE-ALGORITMEN OCH INTILLIGANDE OMRÅDEN

4.3.1 Algoritm utan distansviktning

Resultatet för närmaste granne-algoritmens förmåga att klassificera områden går att se i Tabell 18. Olika radier för buffercirklar och dataunderlag gav upphov till olika resultat. Det är dock främst andelen klassificerade områden som skiljer sig. Om endast 5 % av dataunderlaget har klassificerats på förhand, klassificeras endast ungefär 10% av alla områden om buffercirkeln har en radie på 50 meter. Radien behöver sedan öka till 600 meter för att samtliga områden i Linköping ska klassificeras. Detta eftersom att det inte finns någon garanti för att det finns några redan klassificerade områden i närheten av varje enskilt område. Det framgår dock av resultatet att överenstämmelsegraden är ungefär densamma oberoende av tillgängligt dataunderlag, utan det är främst andelen klassificerade områden som påverkas. Överensstämmelsegraden förefaller dock minska nästintill linjärt i takt med att bufferradien ökar. En större andel av alla områden klassificeras vid samma radie, men bara om den redan klassificerade andelen av områden ökar.

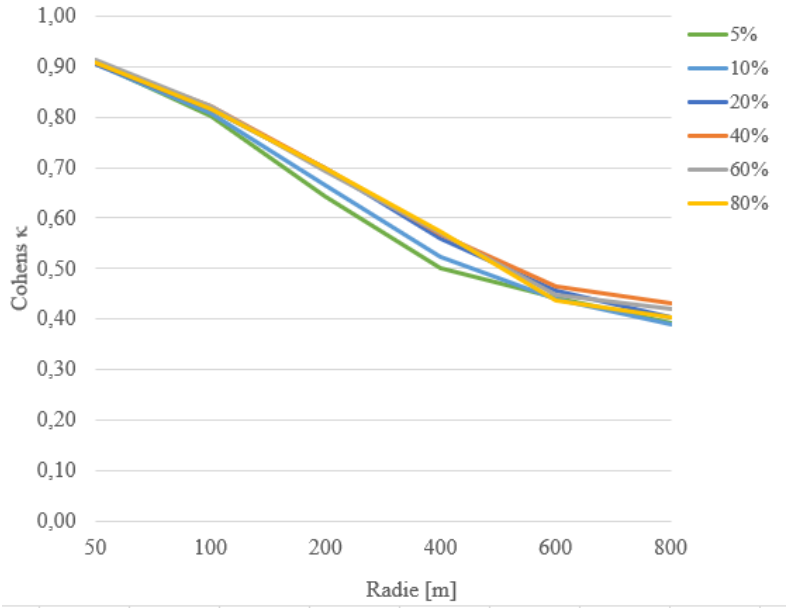
Tabell 18. *Data över närmaste granne-algoritmens prestation, beroende på vald radie för buffercirkel och andel tidigare klassificerade områden. κ-värden över 0,50 är markerade med en asterisk (*).*

Dataunderlag ↓	Bufferrradie $[m]$ →	50	100	200	400	600	800
5%	Noggrannhet	0,95	0,89	0,80	0,74	0,71	0,69
	Avvikelse	0,05	0,11	0,20	0,26	0,29	0,31
	Nollratio	0,44	0,44	0,45	0,47	0,48	0,49
	Cohens κ	0,91*	0,80*	0,64*	0,50	0,44	0,39
	Antal klassade	1105	3445	6565	8204	8393	8456
	Andel klassade	0,13	0,41	0,78	0,97	0,99	1,00
10%	Noggrannhet	0,95	0,89	0,82	0,75	0,71	0,69
	Avvikelse	0,05	0,11	0,18	0,25	0,29	0,31
	Nollratio	0,44	0,44	0,46	0,48	0,48	0,49
	Cohens κ	0,90*	0,81*	0,66*	0,52*	0,44	0,39
	Antal klassade	1835	4917	7327	7893	7984	8015
	Andel klassade	0,23	0,61	0,91	0,98	1,00	1,00
20%	Noggrannhet	0,95	0,90	0,84	0,77	0,72	0,70
	Avvikelse	0,05	0,10	0,16	0,23	0,28	0,30
	Nollratio	0,43	0,43	0,47	0,48	0,49	0,49
	Cohens κ	0,90*	0,82*	0,70*	0,56*	0,46	0,40
	Antal klassade	2783	5683	6847	7103	7103	7112
	Andel klassade	0,39	0,80	0,96	1,00	1,00	1,00
40%	Noggrannhet	0,95	0,90	0,84	0,78	0,73	0,71
	Avvikelse	0,05	0,10	0,16	0,22	0,27	0,29
	Nollratio	0,43	0,45	0,48	0,49	0,49	0,49
	Cohens κ	0,91*	0,82*	0,70*	0,57*	0,46	0,43
	Antal klassade	3177	4890	5312	5382	5391	5395
	Andel klassade	0,59	0,91	0,98	1,00	1,00	1,00
60%	Noggrannhet	0,95	0,90	0,84	0,78	0,72	0,70
	Avvikelse	0,05	0,10	0,16	0,22	0,28	0,30
	Nollratio	0,42	0,46	0,48	0,49	0,49	0,49
	Cohens κ	0,91*	0,82*	0,69*	0,57*	0,45	0,42
	Antal klassade	2591	3419	3621	3652	3657	3657
	Andel klassade	0,71	0,93	0,99	1,00	1,00	1,00
80%	Noggrannhet	0,95	0,90	0,84	0,78	0,71	0,69
	Avvikelse	0,05	0,10	0,16	0,22	0,29	0,31
	Nollratio	0,42	0,46	0,48	0,48	0,48	0,48
	Cohens κ	0,91*	0,82*	0,70*	0,57*	0,44	0,40
	Antal klassade	8604	8867	8901	8971	8986	8990
	Andel klassade	0,96	0,99	0,99	1,00	1,00	1,00

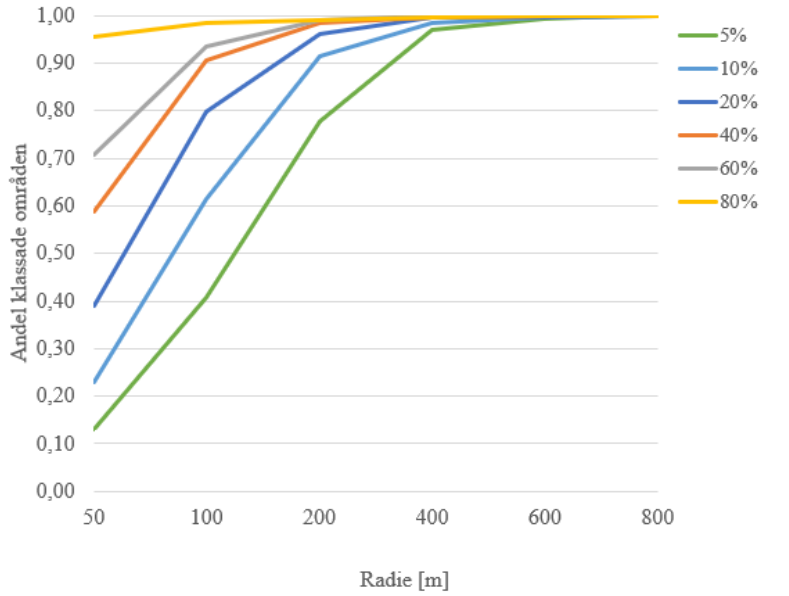
I Figur 26a visas radiens förhållande till Cohens κ, och i Figur 26b har radies förhållande till andelen områden som har klassificerats av algoritmen. Så länge Cohens κ är större

än noll så presterar algoritmen bättre än om den vanligast förekommande områdestypen valts för varje område. Samtliga områden klassificerades när bufferradien utökades till att omfatta alla områden inom 800 meter, och det lägsta κ-värdet var 0,39 för de två lägsta andelarna tillgängligt dataunderlag.

Överensstämmelsegraden motsvarar *Måttligt bra* när den är som sämst vid 800 meter. Däremot uppnår metoden en överenstämmelsegrad som motsvarar *Nästintill perfekt* för alla områden som klassificeras inom en bufferradie av 100 meter.



(a). *Radie mot Cohens κ.*



(b). *Radie mot andelen klassificerade områden.*

Figur 26. *Cohens κ och andelen områden som klassificerats av närmaste granne-algoritmen, beroende på vald radie för buffercirkel. Data är hämtat ur Tabell 18.*

Metoden förefaller att i flera avseenden kunna uppnå ett värde för Cohens κ som motsvarar en överensstämmelsegrad som är nästintill perfekt. Resultat för precision, känslighet och tillhörande F1-värden finns presenterade i Bilaga D.

4.3.2 Algoritm med distansviktning

När metoden med distansviktning tillämpades, kunde samtliga områden klassificeras på en gång. Därför fanns endast behovet av att undersöka hur resultatet påverkades av andelen tillgängligt dataunderlag, alltså andelen sedan tidigare klassificerade områden. Resultaten finns att se i Tabell 19. En *Nästintill perfekt* överensstämmelsegrad uppnås först när 80% av alla områden används för att klassificera resterande. Däremot uppnås en överenstämmelsegrad som är *Bra* redan vid 5% och *Mycket bra* redan vid 10%.

Tabell 19. *Värden för noggrannhet, avvikelse, nollratio och Cohens κ för klassificeringsmetoden baserad på närmaste granne-algoritmen med distansviktning, med 5%, 10% 20%, 40%, 60%, 80% och 100% av tidigare klassificerade områden använda som tillgängligt dataunderlagi Linköping.*

Dataunderlag →	5%	10%	20%	40%	60%	80%
Noggrannhet	0,77	0,82	0,86	0,88	0,90	0,90
Avvikelse	0,23	0,18	0,14	0,12	0,10	0,10
Nollratio	0,49	0,49	0,49	0,49	0,49	0,48
Cohens κ	0,55	0,64	0,72	0,77	0,80	0,81

Metoden erhöll resultat som låg mestadels inom intervallet *Bra-Mycket bra*, och skillnaden mellan att använda till exempel 5% och 10% tillgängligt dataunderlag föreföll inte särskilt stor.

Resultatet för hur metoden presterade för olika bebyggelsetyper finns i Tabell 20. Här framgick skillnaden i prestation beroende på dataunderlaget tydligare. Metoden är tveklöst bättre på att klassificera områden med bebyggelsetyperna F, I, R och V2, än många andra, då dessa erhöll högst F1-värden. Däremot blev resultatet bättre när en större procentandel av områdena gjordes tillgängliga för algoritmen. Det fanns ytterst få områden med bebyggelsetyperna C, FARM, GV och V1 i Linköping i jämförelse med de andra bebyggelsetyperna. Detta återspeglas troligtvis även i resultatet.

Tabell 20. *Data över närmaste granne-algoritmens prestation för olika bebyggelsetyper med tillgång till olika andelar redan klassificerat dataunderlag i Linköping. Värden över 0,50 är markerade med en asterisk (*).*

Data ↓	Bebyggelsetyp →	C	F	FA	GV	I	R	V1	V2
5%	Precision	0,07	0,62*	0,00	0,07	0,84*	0,72*	0,20	0,81*
	Känslighet	0,00	0,47	0,00	0,00	0,60*	0,76*	0,02	0,85*
	F1	0,00	0,53*	0,00	0,00	0,70*	0,74*	0,04	0,83*
10%	Precision	0,07	0,69*	0,00	0,07	0,81*	0,79*	0,12	0,85*
	Känslighet	0,00	0,56*	0,00	0,00	0,67*	0,80*	0,02	0,90*
	F1	0,00	0,62*	0,00	0,00	0,73*	0,79*	0,04	0,87*
20%	Precision	0,22	0,73*	0,86*	0,00	0,90*	0,84*	0,50	0,88*
	Känslighet	0,08	0,62*	0,29	0,00	0,69*	0,86*	0,05	0,93*
	F1	0,11	0,67*	0,43	0,00	0,78*	0,85*	0,09	0,91*
40%	Precision	0,09	0,85*	0,75*	0,00	0,90*	0,87*	0,33	0,90*
	Känslighet	0,05	0,71*	0,43	0,00	0,77*	0,88*	0,06	0,95*
	F1	0,06	0,77*	0,55*	0,00	0,83*	0,88*	0,11	0,92*
60%	Precision	0,33	0,85*	1,00*	0,07	0,90*	0,89*	0,50	0,91*
	Känslighet	0,23	0,71*	0,42	0,00	0,75*	0,89*	0,10	0,96*
	F1	0,27	0,77*	0,59*	0,00	0,82*	0,89*	0,16	0,93*
80%	Precision	0,33	0,84*	1,00*	0,00	0,91*	0,89*	0,50	0,92*
	Känslighet	0,29	0,71*	0,20	0,00	0,75*	0,91*	0,11	0,96*
	F1	0,31	0,77*	0,33	0,00	0,82*	0,90*	0,18	0,94*

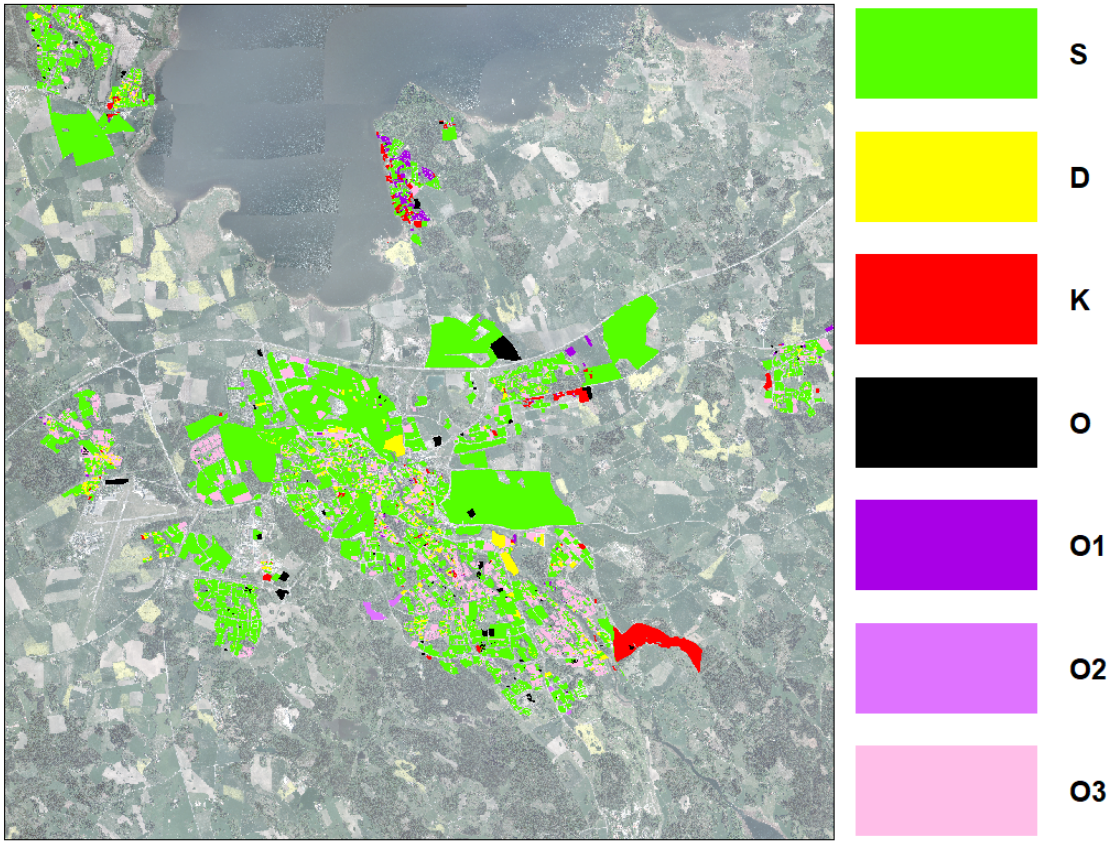
4.4 METOD IV - KLASSIFICERING AV ANSLUTNINGSFÖRHÅLLANDEN

I Figur 27 presenteras klassificeringsresultatet med avseende på anslutningsförhållanden utan servisförlängningar genom att visa fördelningen av områdena på en karta över Linköping. Resultatet som erhölls med förlängning av serviserna går att se i Figur 28. Antalet områden som har klassificerats med en viss typ av anslutningsförhållande går att se i Tabell 21.

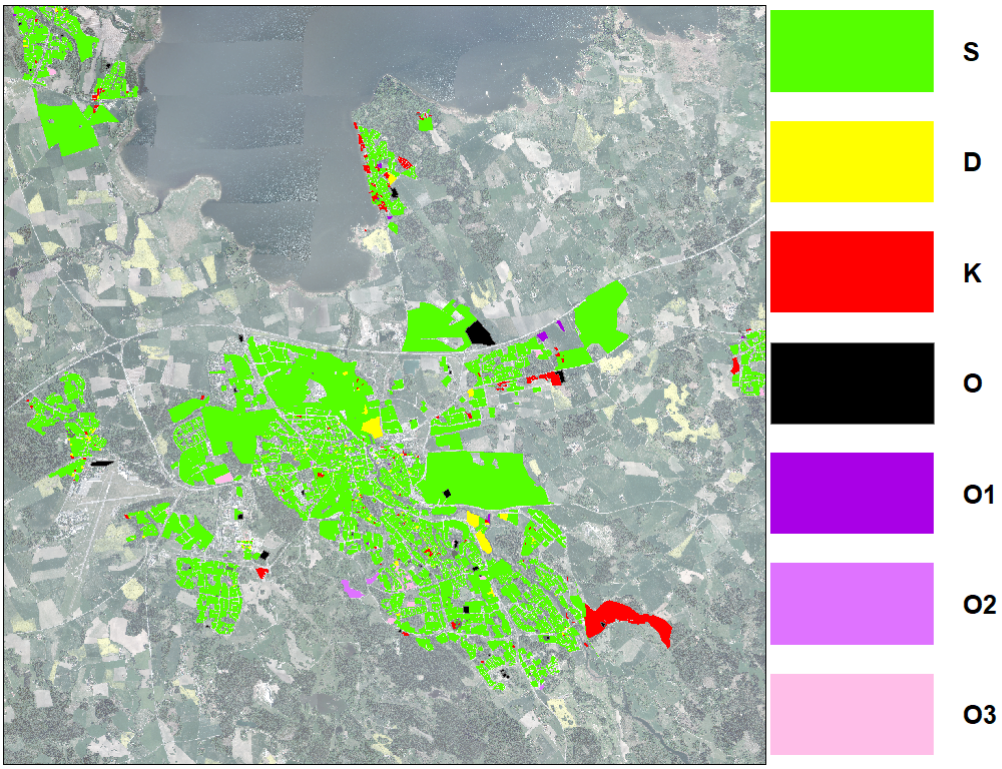
Det är tydligt att majoriteten av områden är separerade. Dock har ett stort antal områden klassificerats som okända, framförallt inom kategorin *O3* som motsvarar områden utan serviser, men med både spill- och dagvattenlednignar i intilliggande gata. Antalet okända områden sjunker drastiskt efter att serviserna har förlängts. Detta bekräftar att ett stort antal områden har serviser som inte är ritade hela vägen fram till områdena. Enstaka kombinerade och delvis kombinerade områden ligger lite utspridda i staden. Områden med separerade ledningssystem är dock fortfarande i majoritet. Vid inspektion av Figur 27 och 28 framgår det dock tydligt att förlängningen av serviser är nödvändig för att en större andel områden ska erhålla korrekt klassificering. Det vore orimligt att ha en så stor andel områden med okända anslutningsförhållanden utan serviser så centralt belägna bland många andra områden med separerade ledningssystem.

Tabell 21. *Antal områden med olika anslutningsförhållanden, klassificerade utan och med datamanipulation.*

Anslutningsförhållande	Antal områden	
	Utan datamanipulation	Med datamanipulation
S	4890	8191
D	655	195
K	200	250
O	214	171
O1	297	33
O2	43	24
O3	2703	138

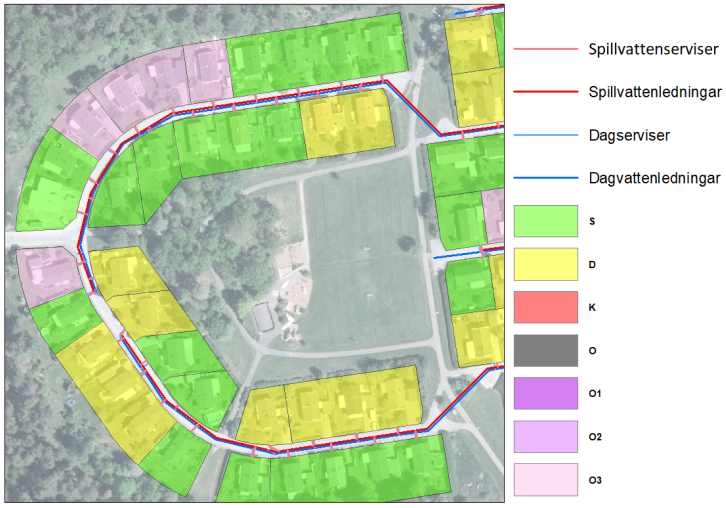


Figur 27. *Klassificeringsresultat med avseende på ledningssystem utan dataunderlagsmanipulation (utan förlängda serviser), där K är kombinerade ledningssystem, D är delvis separerade ledningssystem och S är separerade ledningssystem. O-O3 representerar de områden som ej föll inom någon av kategorierna. ©Lantmäteriet. Bakgrundsbild: Ortofoto, 0,25 m färg (Lantmäteriet, 2019a).*

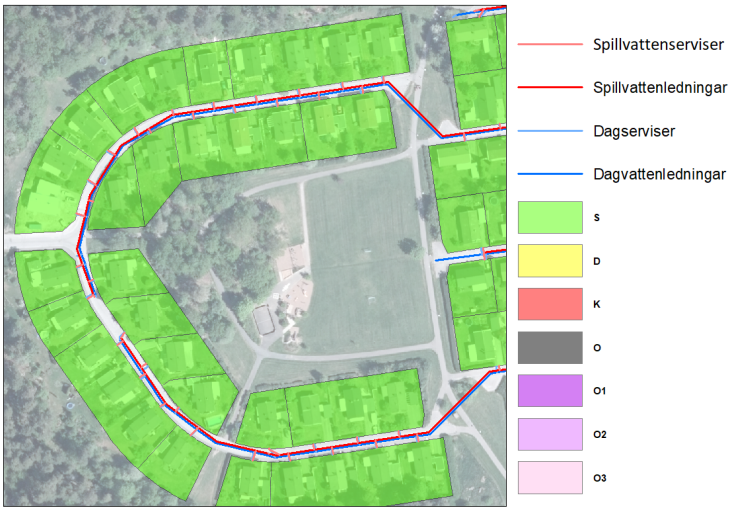


Figur 28. *Klassificeringsresultat med avseende på ledningssystem med dataunderlagsmanipulation (med förlängda serviser), där K är kombinerade ledningssystem, D är delvis separerade ledningssystem och S är separerade ledningssystem. O-O3 representerar de områden som ej föll inom någon av kategorierna ©Lantmäteriet. Bakgrundsbild: Ortofoto, 0,25 m färg (Lantmäteriet, 2019a).*

Genom att undersöka några få områden bekräftades det att många områden som har klassificerats som okända eller delvis kombinerade i själva verket är separerade med avseende på anslutningsförhållanden. Detta berodde på att många serviser eller ledningar inte nådde hela vägen fram till området, så som de är ritade. Ett exempel på detta går att se i Figur 29a. I detta exempel framgår det tydligt att båda spill och dagvattenserviser är anslutna till varje område, ibland flera stycken. Däremot har flera områden trots det klassats klassats som okända för att serviserna inte överlappar med fastigheterna överhuvudtaget, eller så har de klassats som delvis kombinerade för att endast en spillvattenservis när fram till området. Däremot fungerar bufferzonerna precis som avsett. De tillåter områdena att tillgodoräkna sig ledningarna som går i gatan, och därmed klassificeras inga områden som helt okända (*O*). I Figur 29b har förlängningen av serviserna tillämpats, och resultatet såg betydligt bättre ut. Samtliga områden blev korrekt klassificerade.



(a). *Exempel på ett antal områden med separerade ledningssystem felaktigt klassificerade som okända eller delvis kombinerade, utan förlängda serviser.*

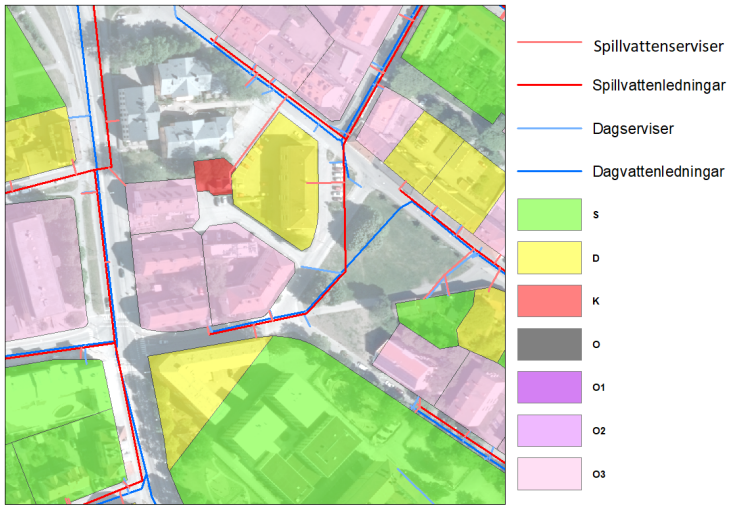


(b). *Exempel på ett antal områden med separerade ledningssystem korrekt klassificerade, med förlängda serviser.*

Figur 29. *Resultatet före och efter förlängningen av serviser, där K är kombinerade ledningssystem, D är delvis separerade ledningssystem och S är separerade ledningssystem. OKÄNT representerar de områden som ej föll inom någon av kategorierna ©Lantmäteriet. Bakgrundsbild: Ortofoto, 0,25 m färg (Lantmäteriet, 2019a)*

Trots att det inte går att utvärdera huruvida metoden har klassificerat alla områden korrekt, förefaller metoden med förlängda serviser mycket lovande vid inspektion av enskilda områden. I Figur 30 syns ytterligare ett exempel på hur metoden presterar med och utan förlängningen av serviser. I Figur 30a har ett flertal områden blivit klassificerade som *O3*. Efter att serviserna har förlängts i Figur 30b så har samtliga områden klassificerats korrekt. Området med kombinerade ledningssystem i mitten är fortfarande korrekt klassificerat eftersom att det bara går en spillvattenservis in till området och det finns ingen intilliggande gata med en dagvattenledning. Även de områden som fortfarande är

delvis kombinerade efter förlängningen i Figur 30b är korrekt klassificerade då dessa saknar dagvattenserviser inom området. Samtliga andra områden som var klassificerade som okända innan förlängningen av serviserna är nu korrekt klassificerade som områden med separerade ledningssystem.



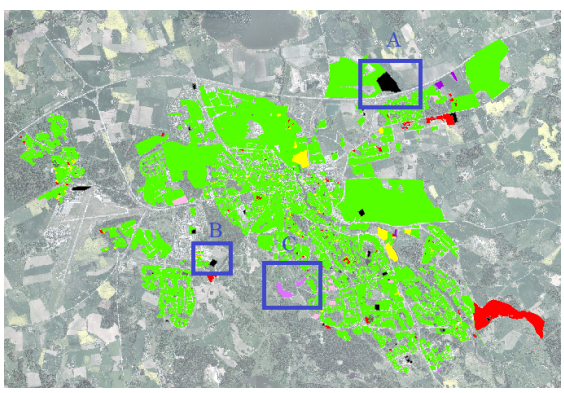
(a). *Exempel på ett antal områden med separerade ledningssystem felaktigt klassificerade som okända eller delvis kombinerade, utan förlängda serviser.*

(b). *Exempel på ett antal områden med separerade ledningssystem korrekt klassificerade, med förlängda serviser.*

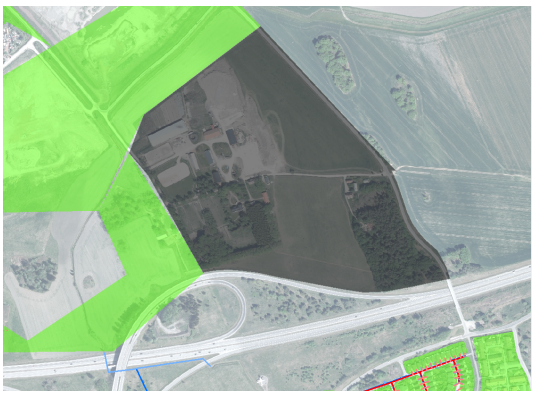
Figur 30. *Resultatet före och efter förlängningen av serviser, där K är kombinerade ledningssystem, D är delvis separerade ledningssystem och S är separerade ledningssystem. OKÄNT representerar de områden som ej föll inom någon av kategorierna ©Lantmäteriet. Bakgrundsbild: Ortofoto, 0,25 m färg (Lantmäteriet, 2019a)*

I Figur 31 visas några exempel på områden som har klassificerats som helt okända (*O*). Mestadels verkar områden som har klassificerats som okända vara områden som ligger i utkant av stadens spill- och dagvattenledningsnätverk. Därför handlar det troligtvis mesta-

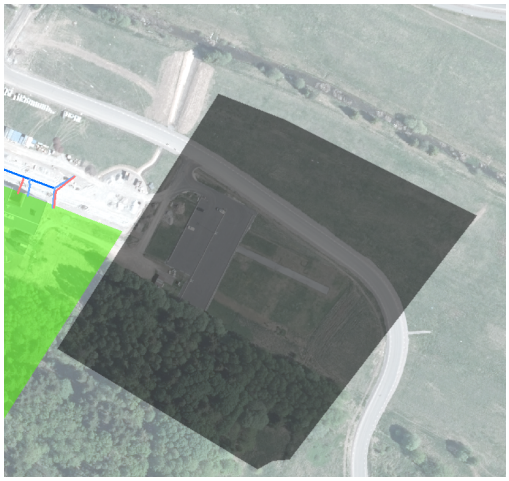
dels om områden som bör erhålla en av de alternativa klassificeringar som omnämns under Avsnitt 2.1.2, delvis obebyggda fastigheter, eller områden där ledningssystemen ej omfattas i dataunderlaget.



(a). *Karta över utvalda och markerade exempelområden med okända anslutningsförhållanden.*



(b). *Exempelområde A.*



(c). *Exempelområde B.*

(d). *Exempelområde C.*

Figur 31. *Ett urval av områden med okända anslutningsförhållanden. ©Lantmäteriet. Bakgrundsbild: Ortofoto, 0,25 m färg (Lantmäteriet, 2019a)*

För att sammanfatta resultatet, förefaller det som att metoden gör det den ska vid en översiktlig inspektion av områden. Det lär dock finnas felaktiga klassificeringar bland de totalt 9002 områden som klassificerades, och utan mer omfattande underlag för utvärdering av metoden är det svårt att garantera eller kvantifiera dess prestation.

4.5 PORTABILITET

4.5.1 Närmaste granne-algoritmen

Närmaste granne-algoritmen var den metoden som gav tveklöst det mest lovande resultatet i jämförelse med samtliga andra metoder för klassificering av områden med avseende

på bebyggelsetyp. Därför testades metoden på ytterligare en plats för att kunna utgöra ett diskussionsunderlag för metodens portabilitet. Portabiliteten för metoden som avser skilja på områden med olika anslutningsförhållanden gick ej att utvärdera, eftersom metoden endast kunde valideras genom inspektion på grund av otillräckligt dataunderlag.

Resultatet för närmaste granne-algoritmen med distansviktning för samtliga områden i Västervik går att se i Tabell 22. Överensstämmelsegraden är ganska mycket lägre än när metoden utvärderas med dataunderlaget från Linköpings kommun. Däremot beror det till stor del på fördelningen av områden av olika bebyggelsetyper. Nollration är i regel ganska mycket större än när metoden tillämpades i Linköping, vilket går att se i Tabell 19. Noggrannheten är fortfarande hög redan vid användande av 5 % tillgängligt dataunderlag. Däremot är κ-värdet ganska lågt vid användande av 5 % tillgängligt dataunderlag, men börjar stiga markant när 40 % eller mer används på grund av att känsligheten ökar.

Tabell 22. *Värden för noggrannhet, avvikelse, nollratio och Cohens κ för klassificeringsmetoden baserad på närmaste granne-algoritmen med distansviktning, med 5%, 10% 20%, 40%, 60%, 80% och 100% av tidigare klassificerade områden använda som tillgängligt dataunderlag i Västervik.*

Dataunderlag →	5%	10%	20%	40%	60%	80%
Noggrannhet	0,67	0,67	0,71	0,80	0,82	0,82
Avvikelse	0,33	0,33	0,29	0,20	0,18	0,18
Nollratio	0,58	0,56	0,58	0,59	0,54	0,51
Cohens κ	0,22	0,22	0,32	0,51	0,62	0,64

Resultatet för hur metoden presterade för olika bebyggelsetyper finns i Tabell 23. Även för Västervik presterar metoden väl för områden av bebyggelsetyperna *F* och *V2*. Däremot inte lika bra för *I* och *R*. Istället presterar metoden mycket bättre i Västervik på områden av bebyggelsetypen *GV*. Detta är rimligt om förekomsten av olika bebyggelsetyper i Västervik jämföres sida vid sida med resultatet. Fördelning av förekomst går att se i Figur 7. Det finns en betydligt större andel klassificerad gatumark och vägyta i jämförelse med områdena som ingick i Linköpings dataunderlag i Tabell 6. Det fanns inte heller lika definierade industri- och radhusområden. Det fanns inte heller lika många kluster som i Linköping.

Tabell 23. *Data över närmaste granne-algoritmens prestation för olika bebyggelsetyper med tillgång till olika andelar redan klassificerat dataunderlag i Västervik. Värden över 0,50 är markerade med en asterisk (*).*

Data ↓	Bebyggelsetyp →	C	F	FA	GV	I	R	V1	V2
5%	Precision	0,07	0,68*	0,07	1,00*	0,50	0,07	0,07	0,61*
	Känslighet	0,07	0,52*	0,07	0,58*	0,06	0,07	0,00	0,97*
	F1	0,07	0,59*	0,07	0,73*	0,11	0,07	0,00	0,75*
10%	Precision	0,07	0,63*	0,07	1,00*	0,50	0,07	0,07	0,63*
	Känslighet	0,07	0,60*	0,07	0,63*	0,06	0,07	0,00	0,91*
	F1	0,07	0,62*	0,07	0,77*	0,11	0,07	0,00	0,74*
20%	Precision	0,07	0,63*	0,07	1,00*	0,36	0,07	0,07	0,71*
	Känslighet	0,07	0,70*	0,07	0,74*	0,16	0,07	0,00	0,87*
	F1	0,07	0,66*	0,07	0,85*	0,22	0,07	0,00	0,78*
40%	Precision	0,07	0,80*	0,07	1,00*	0,56*	0,07	0,07	0,76*
	Känslighet	0,07	0,73*	0,07	0,90*	0,56*	0,07	0,00	0,87*
	F1	0,07	0,76*	0,07	0,95*	0,56*	0,07	0,00	0,81*
60%	Precision	0,07	0,81*	0,07	0,97*	0,62*	0,07	1,00*	0,81*
	Känslighet	0,07	0,74*	0,07	0,93*	0,62*	0,07	0,50*	0,86*
	F1	0,07	0,77*	0,07	0,95*	0,62*	0,07	0,67*	0,83*
80%	Precision	0,07	0,73*	0,07	1,00*	0,67*	0,07	1,00*	0,86*
	Känslighet	0,07	0,89*	0,07	0,85*	0,67*	0,07	0,52*	0,83*
	F1	0,07	0,80*	0,07	0,92*	0,67*	0,07	0,68*	0,85*

5 DISKUSSION

5.1 METOD I - REKTANGULÄRKLASSIFICERING OCH BYGGNADSYTOR

Rektangulärklassificering baserad på de parametrar som testades visade sig vara otillräcklig som metod när det kom till att skilja på områden av olika bebyggelsetyper. Metoden klassificerade endast ungefär två procent av av områdena i Linköping korrekt, vilket är betydligt sämre än om områdena hade klassificerats slumpmässigt. Detta återspeglas av det låga κ-värdet på -1,01.

Resultatet återspeglar dock endast användningen av parametrar baserade på byggnadsytor och områdets area. I teorin skulle det dock vara fullt möjligt att använda metoden om det finns parametrar utan överlappning för de olika bebyggelsetyperna. Metoden är trots allt väldigt enkel - trots att den var komplicerad att tillämpa i FME Workbench. Detta berodde främst på att det innebar en hel del arbete att erhålla minimum- och maximumvärden för samtliga bebyggelsetyper och sedan formulera villkoren för klassificeringsvalen baserat på dessa. Om det skulle vara möjligt att urskilja och definiera klasserna för bebyggelsetyp utifrån ett bestämt antal parametrar skulle det resultera i en mer konsekvent klassificeringsmetod. Detta eftersom att den mänskliga faktorn minimeras när områdena klassificeras helt och hållet baserat på ett antal urvalskriterier.

Det är även möjligt att använda en kombination av flera olika typer av parametrar för att skilja på enskilda bebyggelsetyper. Till exempel, skulle det eventuellt gå att hitta en parameter som skiljer en klass från andra klasser. De något högre *F1*-värdena för bebyggelsetyperna *F* (flerfamiljshus) och *GV* (gatu- och vägyta) tyder på att dessa skiljer sig från andra bebyggelsetyper mer än andra. Resultatet är dock fortfarande inte tillräckligt bra för att det ska vara användbart i en klassificeringsmetod. Däremot tyder det på att det är möjligt att finna en bättre kombination av urvalsvillkor. Framförallt om fler parametrar kunde undersökas. Det skulle dock bli tidskrävande och endast nödvändigt om behovet för ett nytt ramverk skulle vara önskvärt.

5.2 METOD II - KLASSIFICERING BASERAD PÅ STÖRSTA SANNOLIKHET

Klassificeringsmetoden baserad på största sannolikhet utvecklades i hopp om att komplettera metoden som utgick från rektangulärklassificering. Resultatet blev betydligt bättre med metoden, men κ-värdet låg fortfarande under noll och metoden lyckades aldrig klassificera mer än 35% av områdena korrekt. Alltså presterade metoden sämre än om den vanligaste bebyggelsetypen hade valts för varje klassificering. Det är dock viktigt att ta hänsyn till att nollration var ganska hög med ett värde på 0,48, vilket indikerar att 48% av områdena hade klassificerats korrekt om den vanligaste bebyggelsetypen (*V2*) hade valts för varje område.

En positiv möjlighet med resultatet är att det visar på att gränser för hur mycket träningsdata som behövs ligger lägre än vad som undersöktes. Resultatet blev aldrig bättre när mer träningsdata användes för att skapa signaturfilerna. Det innebär att om metoden

skulle tillämpas så är det överflödigt att klassificera så mycket som 20% av områdena som träningsdata. Å andra sidan, så skulle resultatet även kunna innebära att det inte finns någon särskild undre eller övre gräns för att parametrarna helt enkelt är otillräckliga för att kunna skilja på områdena överhuvudtaget. Vidare, om endast ett fåtal områden används för att definiera klasserna så bör dessa vara väl bedömda typområden - annars lär metoden prestera bättre med fler. Dataunderlaget från Linköping bestod av 9002 områden, därför motsvarar 20% av dataunderlaget 1800 områden. Det är fullt möjligt att det hade räckt att klassificera några enstaka områden av varje bebyggelsetyp för att erhålla samma resultat med ett lågt κ-värde.

När det kommer till att skilja enskilda bebyggelsetyper från andra visar dock metoden lovande resultat för *V2*-områden. Redan vid 20% tillgängligt dataunderlag är $F1$-värdet 0,59. Det råder en jämn balans mellan precision och känslighet - alltså är resultatet troligtvis inte slumpmässigt. Därmed klassificerar metoden 64% av alla *V2*-områden korrekt, åtminstone i Linköping. För att utvärdera metoden hade det varit nödvändigt att tillämpa den på flera städer. Vidare, vore det även nödvändigt att utreda om samma signaturfiler kan användas på flera städer eller om det är nödvändigt att skapa träningsdata där den ska appliceras. Detta hade troligtvis gett upphov till varierande resultat eftersom att områden av olika bebyggelsetyper kan skilja sig med avseende på olika parametrar från stad till stad.

Varken de parametrar som undersöktes baserade på byggnadsytor eller servisförekomst resulterade i en prestation som gör metoden användbar i dagsläget. Även för denna metod finns dock möjligheten att vidare undersöka huruvida det finns parametrar som inte överlappar varandra till den grad att de inte går att skilja från varandra. Det är även teoretiskt sett en bättre metod än rektangulärklassificering, då en del överlappning mellan parametrar oftast är hanterbart. Det finns även redan färdigutvecklade verktyg i ArcGIS för att utföra klassificeringen som gör metoden mycket mindre tidskrävande. Vidare, finns det även möjligheter att vikta de olika bebyggelsetyperna. Därmed har metoden ett flertal optimeringsmöjligheter.

5.3 METOD III - NÄRMASTE GRANNE-ALGORITMEN OCH INTILLIGGANDE OMRÅDEN

Metoden som baserats på närmaste granne-algoritmen är tveklöst den metod som har uppvisat mest lovande resultat. Av den anledningen var det även endast för denna metoden som portabilitet undersöktes. Resultatet för testet i Västervik är inte tillräckligt för att kvantifiera hur väl metoden kan tänkas prestera, men utgör åtminstone ett underlag för spekulation.

När olika sökradier för närmaste granne-algoritmen testades förefaller förhållandet mellan överensstämmelsegraden och buffercirklarnas radie vara nästintill linjärt. Detta framgår i Figur 26a. Detta stämmer överens med Toblers lag som dikterar att allting är relaterat till allting annat, men att näraliggande enheter relaterar mer till varandra i jämförelse med andra enheter som ligger längre bort (Tobler, 1970). Beroende på var metoden tillämpas

så kommer dock överenstämmelsegraden att variera. Detta framgår när resultaten som erhölls med dataunderlaget från Linköping jämförs med vad som erhölls med hjälp av dataunderlaget från Västervik. Även för dessa dataunderlag måste värdet för nollration tas hänsyn till. En stor andel av områden hade klassificerats korrekt om den vanligaste bebyggelsetypen hade valts för varje område. Ett tillfredsställande resultat erhölls dock med endast en liten andel av det tillgängliga dataunderlaget, för att sedan bli bättre och bättre ju större den andelen blev. Däremot verkar κ-värdet uppnå ett maximum kring 0,90 för dataunderlaget från Linköping, och kring 0,60 för dataunderlaget från Västervik. Det framgår även att metoden presterar väl för de områden som är i majoritet, vilket är fullt rimligt eftersom att områdena klassificeras baserat på andra intilliggande områden. Det finns en del skillnader mellan de olika städerna och förekomsten av olika bebyggelsetyper. Detta återspeglas i resultaten, då de högsta *F1*-värdena varierar mellan de olika städerna. Följaktligen bör stadens utformning tas hänsyn till vid eventuellt användande av metoden. Ju mer slumpmässigt och blandat områden av olika bebyggelsetyper är utplacerade, ju sämre presterar metoden. Är staden däremot planerad på så sätt att industri-, villa- och radhusområden är åtskiljda så finns det goda chanser för att metoden ger upphov till ett bättre resultat. Av logistiska skäl är detta ofta fallet. Resultatet för de enskilda klasserna återspeglar även hur viktigt det är att ta hänsyn till huruvida enskilda bebyggelsetyper kan tänkas vara placerade i kluster. De klasser med högst *F1*-värden är *V2*- och *I*- områden. Alltså villaområden med tomter större än 1000 m^2 och industriområden. Dessa är typiska områden som kan tänkas vara geografiskt grupperade.

En fördel med den utvecklade metoden som omfattar en slags distansviktning är att den alltid klassificar alla områden baserat på närmast möjliga intilliggande områden. Däremot finns det utrymme för vidare optimering av använda radier, då dessa kan tänkas variera något för olika städer. Å andra sidan kommer överenstämmelsegraden troligen att minska med radiernas storlek och skillnaderna som andra storlekar för radierna ger upphov till kan bli obetydliga. En annan möjlighet vore att enskilt distansvikta varje områdes relation till området som ska klassificeras. Den främsta anledningen till att det är en fördel att använda cirklar istället för att enskilt distansvikta varje område är att metoden skulle bli väldigt beräkningstung annars.

Beroende på noggrannhetskrav och vad det är som ska klassificeras har metoden flera användningsområden. Förutsatt att övervägda klasser kan tänkas vara grupperade på ett eller annat sätt har metoden goda möjligheter till att utgöra ett användbart klassificeringsverktyg. Själva algoritmen är trots allt enkel. Beroende på tidigare erfarenhet kan det vara tidskrävande att utforma metoden i ett program som FME. Fördelen är att metoden kan användas med olika dataunderlag utan att några förändringar är nödvändiga.

5.4 METOD IV - KLASSIFICERING AV ANSLUTNINGSFÖRHÅLLANDEN

När det kommer till att skilja på områden med olika anslutningsförhållanden så fanns det inget konkret sätt att utvärdera metoden. Detta eftersom att olika delar av dataunderlaget inte var komplett både i Linköping och Västervik. I dataunderlaget från Linköping saknades information om anslutningsförhållanden och i dataunderlaget från Västervik så

saknades dag- och spillvattenledningar. Därför var det inte möjligt att utvärdera klassificeringsmetoden med förväxlingsmatriser och erhålla värden för noggrannhet och överensstämmelsegrad. Istället gjorde en visuell rimlighetsbedömning och ett fåtal enskilda områden inspekterades i syfte att bättre förstå hur metoden presterade.

Ett vanligt problem tycks vara att serviser inte når hela vägen fram till fastigheterna så som de är utritade (Hammarlund, 2019). Detta medför problem när anslutningsförhållanden avgörs automatiskt av en dator eftersom det blir svårt att formulera villkor som utgår från förekomsten av olika typer av serviser och ledningar. Detta framgick även av resultaten. Många områden som bör ha klassificerats som områden med separerade ledningssystem klassificerades istället som delvis kombinerade eller okända eftersom att serviserna avsedda för områdena ofta inte sträckte sig hela vägen fram till områdena i fråga. Detta adresserades genom att manipulera dataunderlaget i metoden. Serviserna förlängdes fem meter och det verkade förbättra resultatet eftersom att många områden som felaktigt klassificerats som områden med kombinerade ledningssystem istället klassificerades som områden med separerade ledningssystem när samtliga serviser nådde fram till området. Det var dock inte möjligt att optimera servislängden eftersom att ett underlag för utvärdering saknades. Däremot verkade det som att det uppstår problem om serviser förlängs ytterligare på grund av att de i tätbebyggda områden kan sträcka in på områden för vilka de inte är avsedda.

I enstaka fall kan serviserna ligga i en vinkel mot ett hörn av området. Detta innebär att serviser inte kommer att överlappa inom området trots att den förlängs. Däremot verkar detta vara ovanligt. Att använda bufferzoner kring ett sådant område i syfte att de även skulle kunna tillgodoräkna sig serviserna, även om dessa ligger utanför området skulle vara ett sätt att lösa problemet. Däremot uppstår problem eftersom att det även finns en möjlighet att området överlappar med serviser ej avsedda för just det området. Därför är det rimligt att sträva efter att så få områden som möjligt klassificeras med ett sådant tillägg i metoden. Därför skulle det även gå att argumentera för att bufferzonerna i det syftet är överflödiga och att enstaka felaktiga klassificeringar vore att föredra eftersom att det inte finns någon garanti för att klassificeringarna blir korrekta med bufferzonerna.

Däremot förefaller det som om bufferzonerna fungerar precis som tänkt när det kommer till att para ihop områden med ledningar i intilliggande gator. Det finns givetvis en risk för att ett område skulle kunna tillgodogöra sig en ledning som egentligen inte hör till området. Just av den anledningen begränsades bufferzonerna till att sträcka sig 15 meter utanför områdena. Om de är större än så blir risken större att området tillgodoräknar sig ledningar som inte hör till området i fråga. Däremot kan de inte heller bli så mycket mindre eftersom att de fortfarande måste sträcka sig så pass långt utanför området att de kan tillgodoräkna sig ledningar som ligger i en intilliggande gata. Det finns en viss risk att ett område kan klassificeras som kombinerat när en dagvattenledning ligger längre bort än tio meter från områdets gräns. Däremot minskar sannolikheten att ett områdes eventuella dagvattenserviser är anslutna till en dagvattenledning ju längre bort den ligger från området. Följaktligen blir den kombinerade klassificeringen acceptabel när metoden används i syfte att förenkla avrinningsberäkningar.

Områden med okända anslutningsförhållanden tycktes främst ligga i utkanten av stadens ledningsnätverk. Dessa har troligtvis alternativa anslutningsförhållanden, vilka är omnämnda under Avsnitt 2.1.3. Antingen det, eller så är dataunderlaget helt enkelt otillräckligt för att utföra klassificeringen. Det kan även handla om obebgyggda fastigheter. Detta är troligtvis det viktigaste att ta hänsyn till när det kommer till just den här metoden. Överenstämmelsegraden kommer att bero helt och hållet på dataunderlaget. Beroende på hur dataunderlaget ser ut så finns alltid möjligheten att manipulera och ändra det i viss utsträckning. Däremot blir metoden något mer svårtillämpad om den ska användas i flera olika städer - framförallt utan utvärderingsmöjligheter. Det positiva med metoden är att den utgår från väldigt strikta klassificeringsvillkor. Den klassificerar helt och hållet beroende på ledning- och servisförekomst. Metoden är inte heller särskilt tidskrävande att utforma i ett program som FME. Därmed kan den endast förbättras genom identifierandet av scenarion där dataunderlaget kan tänkas vara begränsande. Metoden förutsätter även att avgränsingen av områden har skett på ett sätt som utelämnar gator och vägyta eftersom att bufferzonerna används för att tilgodoräkna områdena korrekta spill- och dagvattenledningar.

Sammanfattningsvis, skulle metoden mycket väl kunna användas i syfte att göra överslags-beräkningar för avrinning. Tidigare studier har visat på att kunskapen om områdens anslutningsförhållande tillsammans med kunskapen om områdens bebyggelsetyp möjliggör noggrannare beräkningar för avrinning (Larsson, 2010). Metoden gör det den ska, men det finns ingen som helst garanti för att underlaget som används är tillräckligt bra för att alla områden ska klassificeras korrekt. Därmed bör den användas med försiktighet, och klassificerade områden bör åtminstone inspekteras översiktligt. Vid krav på större noggrannhet är troligtvis en mer noggrann visuell inspektion nödvändig. Hur många områden som inspekteras är avgörande när det kommer till hur tidskrävande metoden är. Däremot tar det inte särskilt lång tid att utföra klassificeringarna eller att utforma metoden.

5.5 VIDARE STUDIER

Även om ett antal klassificeringsmetoder har jämförts så har denna studie knappt skrapat på ytan. Det finns ett stort antal metoder för klassificering som har utvecklats i olika syften. Utmaningen består i att anpassa metoderna för klassificering av område med avseende på bebyggelsetyp eller anslutningsförhållanden.

Metoden som klassificerar baserad på största sannolikhet har goda möjligheter att prestera väl med tanke på hur den är uppbyggd. Överlappning inom vissa parametrar är förlåtligt om det finns fler parametrar där överlappningen inte är lika stor. För att få den att fungera skulle fler parametrar behöva testas. Till exempel skulle det kunna vara en möjlighet att använda metoden som det ursprungligen är tänkt för att bestämma landklasser som olika typer av hårdlagda ytor och grönyta, för att sedan använda förhållanden mellan dessa för att skapa nya parametrar vilka i sin tur skulle kunna användas för att skilja på olika bebyggelsetyper.

Det hade även varit intressant att undersöka om den metod som är baserad på närmaste granne-algoritmen skulle kunna kompletteras med en annan metod. Om till exempel en metod skulle kunna användas för att med en större säkerhet klassificera det underlaget som krävs för att använda närmaste granne-algoritmen så skulle det spara ännu mer tid. Vidare, borde metoden testas på fler dataunderlag från olika städer i syfte att utvärdera hur stadens utformning, samt områdenas avgränsning och fördelning påverkar metodens överensstämmelsegrad.

Slutligen, så skulle även metoden för klassificering av områden med avseende på anslutningsförhållanden kunna både utvärderas, samt utvecklas ytterligare. Det finns en viss andel områden som klassificeras som okända. De områden som erhållit dessa klassificeringar skulle behöva undersökas i större detalj för att avgöra hur eventuella felklassificeringar kan undvikas och hur dessa bör hanteras. Det hade även varit användbart att ta fram riktlinjer för hur serviser och ledningar bör manipuleras beroende på metoden implementeras. Återigen begränsas metoden av var den implementeras och det tillgängliga dataunderlaget.

6 SLUTSATSER

I examensarbetet utvecklades tre alternativa metoder för områdesklassificering med avseende på bebyggelsetyp, samt en metod för områdesklassificering med avseende på anslutningsförhållanden. Frågeställningarna kan besvaras enligt följande punkter.

- Det gick att utveckla en fungerande alternativ metod för områdesklassificering med avseende på bebyggelsetyp. En metod baserad på utveckling av närmaste granne-algoritmen uppvisade lovande resultat i överensstämmelsegrad både i Linköping och Västervik.
- En metod för klassificering av områden med avseende på anslutningsförhållanden utvecklades framgångsrikt. Vid översiktlig visuell inspektion förefaller modellen prestera väl, vilket innebär att den eventuellt skulle kunna användas med försiktighet.
- Metodvalet har en stor betydelse för klassificering av områden. Rektangulärklassificering förefaller otillräcklig då parametrar för områden av olika bebyggelsetyper tenderar att överlappa. Klassificering baserad på största sannolikhet erhåller ett bättre resultat i jämförelse, men resultatet är fortfarande inte tillräckligt bra för att metoden ska kunna tillämpas praktiskt.
- Vilket underlag som krävs beror helt och hållet på metodval och följaktligen parameterval. Inget underlag utöver vanligt förekommande kartmaterial tillgängligt via kommuner och Lantmäteriet användes för metodutvecklingen.
- Metoden för klassificering av områden med avseende på anslutningsförhållanden är enklast och tog minst tid att utveckla. De andra metoderna för klassificering av områden med avseende på bebyggelsetyp är alla något mer tidskrävande. Metoden som är baserad på närmaste granne-algoritmen, vilken presterade bäst, är tidskrävande i den aspekten att ett visst antal områden med jämn geografisk spridning behöver klassificeras på förhand. Tidsåtgång kommer även att variera beroende på tidigare användarerfarenhet av använd programvara.

Metoden som utvecklades för områdesklassificering av bebyggelsetyper begränsas av städers utformning samt hur områden har avgränsats och är fördelade. Därmed kräver användandet av metoden en förståelse för hur metoden fungerar och under vilka förutsättningar den kan tänkas prestera sämre. Metoden för områdesklassificering av anslutningsförhållanden skulle behöva ett mer omfattande dataunderlag från fler städer i syfte att utvärdera metoden och avgöra hur väl den presterar i olika städer med annorlunda utformning och dataunderlag. Tillgängligt dataunderlag kan ge upphov till en del begränsningar. Om delar av ledningsnätverket saknas eller är bristfälligt återgivet i vektorform kommer det att påverka de klassificeringar som erhålles.

7 REFERENSER

Ahmad, A. och Quegan, S. (2012). *Analysis of Maximum Likelihood Classification on Multispectral Data.* Applied Mathematical Sciences, 6(129), s. 6425-6436.

Arnell, V. (1980). *Dimensionering och analys av dagvattensystem - val av beräkningsmetod.* Göteborg: Geohydrologiska forskningsgruppen, Chalmers tekniska högskola.

Blomquist, D., Hammarlund, H., Härle, P. och Karlsson, S. (2016). *Riktlinjer för modellering av spillvattenförande system och dagvattensystem.* Svenskt Vatten, 2016-15.

Bolstad, P. V. och Lillesand, T. M. (1991). *Rapid Maximum Likelihood Classification.* Photogrammetric Engineering and Remote Sensing, 57(1), s. 67-74.

Cohen, J. (1960). *A coefficient of agreement for nominal scales.* Educational and Psychological Measurement, XX(1).

ESRI (2019a). *How Maximum Classification Works*. URL: http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-maximum-likelihood-classification-works.htm (hämtad 2019-10-29).

ESRI (2019b). *Om GIS*. URL: http://www.esri.se/om-gis (hämtad 2019-10-01).

ESRI (2019c). *Overview of Image Classification*. URL: https://pro.arcgis.com/en/pro-app/help/analysis/image-analyst/overview-of-image-classification.htm (hämtad 2019-10-29).

Hwang, W. J. och Wen, K. W. (1998). *Fast KNN classification algorithm based on partial distance search.* Electronics Letters, 34(21), s. 2062-2063.

Hammarlund, H. (2019). *Specialist i Hydraulisk modellering, Tyréns AB, Möte 2019-11-29.*

Helsel, D. R. och Hirsch, R. M. (2002). *Statistical Methods in Water Resources*. Techniques of Water-Resources Investigations of the United States Geological Survey, Book 4, Hydrologic Analysis and Interpretation, s. 524.

Kataria, A. och Singh, M. D. (1967). *Nearest neighbor pattern classification.* IEEE Trans. Information Theory, 13, s. 21-27.

Kataria, A. och Singh, M. D. (2013). *A Review of Data Classification Using K-Nearest Neighbour Algorithm.* International Journal of Emerging Technology and Advanced Engineering, ISO 9001:2008 Certified Journal, 3(6).

Koch, G.G. och Landis, R. (1977). *The Measurement of Observer Agreement for Categorical Data.* Biometrics, 33(1), s. 159-174.

Lantmäteriet (2019a). *Ortofoto över Linköping, 0,25 m färg [Kartografiskt material]*. URL: https://www.lantmateriet.se/sv/Kartor-och-geografisk-information/geodataprodukter/ortofoto/ (hämtad 2019-09-08).

Lantmäteriet (2019b). *Ortofoto över Västervik, 0,25 m färg [Kartografiskt material]*. URL: https://www.lantmateriet.se/sv/Kartor-och-geografisk-information/geodataprodukter/ortofoto/ (hämtad 2019-11-28).

Larsson, J. (2010). *Metodik för beräkning av anslutna hårdgjorda ytor till spillvattennätet.* Examensarbete, Uppsala universitet.

Matasci, G., Longbotham, N., Pacifici, F., Kanevski, M. och Tuia, D. (2019). *Understanding angular effects in VHR imagery and their significance for urban lan-cover model portability: A study of two multi-angle in-track image sequences*. ISPRS Journal of Photogrammetry and Remote Sensing, 107, s. 99-111.

Nationalencyklopedin (2019a). *Avlopp*. URL: https://www-ne-se.ezproxy.its.uu.se/uppslagsverk/encyklopedi/l%C3%A5ng/avlopp (hämtad 2019-09-21).

Nationalencyklopedin (2019b). *Dagvatten*. URL: https://www-ne-se.ezproxy.its.uu.se/uppslagsverk/encyklopedi/l%C3%A5ng/dagvatten (hämtad 2019-09-21).

Nationalencyklopedin (2019c). *Hydrologi*. URL: https://www-ne-se.ezproxy.its.uu.se/uppslagsverk/encyklopedi/l%C3%A5ng/hydrologi (hämtad 2019-09-21).

Nationalencyklopedin (2019d). *Täthetsfunktion*. URL: https://www-ne-se.ezproxy.its.uu.se/uppslagsverk/encyklopedi/l%C3%A5ng/t%C3%A4thetsfunktion (hämtad 2019-11-28).

Perumal, K. och Bhaskaran, R. (2010). *Supervised Classification Performance of Multispectral Images.* Journal of Computing, 2(2).

RStudio (2019). *About Rstudio*. URL: https://rstudio.com/about/ (hämtad 2019-10-23).

Safe Software (2019a). *FME*. URL: https://www.safe.com/fme/ (hämtad 2019-10-03).

Safe Software (2019b). *FME Transformer Gallery*. URL: https://www.safe.com/transformers/ (hämtad 2019-10-10).

Shapiro, S. S., Wilk, M. B. och Chen, H. J. (1968). *A Comparative Study of Various Tests for Normality.* Journal of the American Statistical Association, 63(324), s. 1343-372.

Sivapalan, M., Troch, P.A., Wagener, T. och Woods, R. (2007). *Catchment classification and Hydrologic similarity.* Geography Compass, 1/4 (2007), s. 901-931.

SMHI (2017). *Avrinning*. URL: https://www.smhi.se/kunskapsbanken/hydrologi/avrinning-1.110938 (hämtad 2019-09-20).

SMHI (2018). *Avrinningsområde*. URL: https://www.smhi.se/kunskapsbanken/hydrologi/avrinningsomrade-1.6704 (hämtad 2019-09-09).

Sokolova, M., Japkowicz, N. och Szpakowicz, S. (2006). *Beyond Accuracy, F-Score and ROC: a Family of Discriminant Measures for Performance Evaluation.* Australasian joint conference on artificial intelligence, s. 1015-1021, Springer, Berlin, Heidelberg.

Svenskt Vatten (2010). *Avledning av dag-, och drän- och spillvatten, Funktionskrav, Hydraulisk dimensionering och utformning av allmänna avloppssytem.* Publikation P110.

Tallón-Ballesteros, A.J. och Riquelme, J. (2014). *Data Mining Methods Applied to a Digital Forensics Task for Supervised Machine Learning.* Studies in Computational Intelligence, 555, s. 413-428.

TechTarget (2019). *Portability*. URL: https://searchstorage.techtarget.com/definition/portability (hämtad 2019-12-09).

The R Project for Statistical Computing (2019). *What is R?* URL: https://www.r-project.org/about.html (hämtad 2019-10-23).

Ting, K.M. (2017). *Confusion Matrix.* In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning and Data Mining. Springer, Boston, Massachusetts.

Tobler, W.R. (1970). *A Computer Movie Simulating Urban Growth in The Detroit Region.* Economic Geography, Vol. 46, Supplement: Proceedings. International Geographical Union. Commision on Quantitative Methods (Jun., 1970). s. 234-240, Clark University.

Ukrainski, P. (2017). *Supervised Image Classification Using Parallelepiped Algorithm*. URL: http://www.50northspatial.org/supervised-image-classification-using-parallelepiped-algorithm/ (hämtad 2019-10-29).

Visa, S., Ramsay B. Ralescu, A. och van der Knaap, E. (2011). *Confusion Matrix-based Feature Selection.* Proceedings of the Twenty-second Midwest Artificial Intelligence and Cognitive Science Conference, University of Cincinnati. Omnipress, Madison, Wisconsin.

BILAGOR

BILAGA A - TRANSFORMATORER I FME DESKTOP

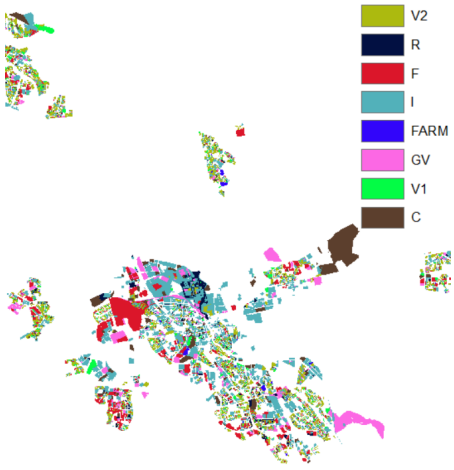
Tabell A.1. *Använda transformatorer i FME desktop och deras beskrivningar (Safe Software, 2019b).*

Transformator	**Beskrivning**
Aggregator	Kombinerar olika attribut och geometri till geografiska enheter, vilka även kan grupperas utifrån attribut eller andra egenskaper.
Arcstroker	Konverterar bågar till linjer genom interpolering.
AreaCalculator	Beräknar arean av en polygon och lagrar i attributform.
Attributemanager	Genomför förändringar på enheter genom att lägga till, ändra namn, ta bort, eller att omstrukturera dem. Det går även att bestämma värden för nya eller existerande attribut med villkorsbestämning eller konstanter.
CenterPointReplacer	Ersätter enhetens geometri med en punkt, antingen i enhetens geometriska centrum, masscentrum eller vid en annan definierad plats inom enhetens area.
Clipper	Klipper och grupperar enheter inom gränserna för andra enheter.
Duplicatefilter	Detekterar dubletter med avseende på attribut.
ExpressionEvaluator	Utför en matematisk beräkning på ett uttryck bestående av bland annat enhetsfunktioner ur FME eller matematiska funktioner.
FeatureJoiner	Kombinerar enheters attribut och/eller geometri.
GeographicBufferer	Skapar en bufferzon av specifierad storlek och form omkring eller inuti en enhets geometri.
GeometryFilter	Sorterar enheter baserat på geometrityp.
GeometryValidator	Detekterar problem hos geometriska enheter och reparerar dem.
ListExploder	Extraherar alla enheters listor, samt gör dessa till egna enheter.
PointOnAreaOverlayer	Undersöker överlappning för punkter inom polygoner, samt assignerar dessa punkter motsvarande polygoners attribut.
PointOnRasterValueExtractor	Extraherar bandvärden från ett raster vid en eller flera punkter, och lagrar dessa som attribut.
RandomNumberGenerator	Genererar randomiserade jämnt fördelade tal.
RasterCellCoercer	Dekomposerar ett raster till individuella punkter eller polygoner.
StatisticsCalculator	Beräknar ett flertal statistika mått baserade på bestämda attribut. Det går även att gruppera enheter och beräkna statistik utifrån enskilda eller fler attribut tillsammans.
Sorter	Sorterar enheter efter ett specifierat attribut, i en given ordning - till exempel alfabetisk eller numerisk.
SpatialFilter	Filtrerar enheter som punkter, linjer, areor eller text baserat på rumsliga förhållanden. Varje kandiderande enhet jämföres mot filter som utgörs av andra enheter.
Tester	Låter varje enhet genomgå ett eller flera test, och låter grupperar dem sedan utefter huruvida testvillkoren uppfylldes eller inte.
Testfilter	Filtrerar enheter med hjälp av ett set av testvillkor och dirigerar dem sedan i grupper till en eller flera utkomstportar beroende på utfall.

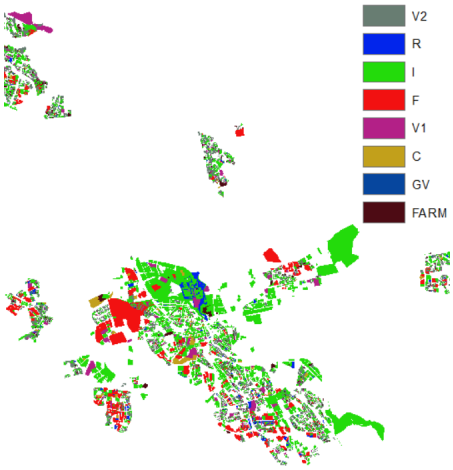
BILAGA B - REKTANGULÄRKLASSIFICERING BASERAD PÅ BYGGNADSYTOR

Tabell B.1. *Statistik använd för klassificeringsvillkor baserad på dataunderlaget från Linköping.*

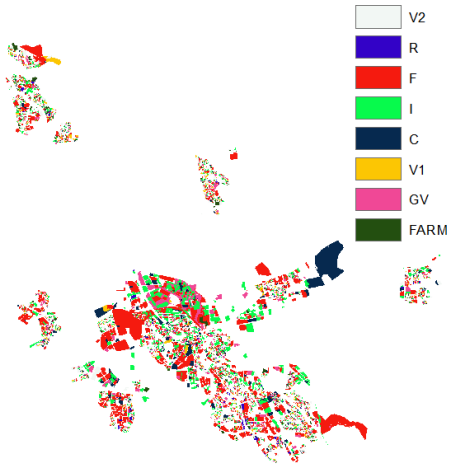
	C	F	FARM	GV	I	R	V1	V2
Min summa [m^2]	7	6	14	9	7	7	20	2
Max summa [m^2]	17352	47026	178860	44484	32021	10636	6000	178860
Median summa [m^2]	230	2062	708	1479	2234	376	328	229
Medel summa [m^2]	1393	3129	7852	4762	3730	624	727	321
STD_sam summa [m^2]	4139	3619	32979	11342	4629	873	1098	2638
STD_pop summa [m^2]	4076	3615	32406	10958	4624	874	1087	2638
Min byggnadsyta [m^2]	1	0	4	1	0	0	2	1
Max Byggnadsyta [m^2]	17115	17289	7848	14190	32014	7096	2851	5321
Median byggnadsyta [m^2]	42	74	153	164	408	63	59	66
Medel byggnadsyta [m^2]	403	278	449	674	1155	98	127	76
STD_sam byggnadsyta [m^2]	2248	617	903	1543	2377	192	308	72
STD_pop byggnadsyta [m^2]	2238	617	900	1536	2376	192	307	72
Min ant. fast. [n]	1	1	1	1	1	1	1	1
Max ant. fast. [n]	15	135	96	65	28	102	21	96
Median ant. fast. [n]	2	4	2	1	2	4	4	3
Tot ant. fast. [n]	114	5812	199	106	1521	20530	269	17198
Medel ant. fast. [n]	3	11	7	7	3	6	6	4
STD_sam ant. fast. [n]	4	20	18	16	3	8	4	3
STD_pop ant. fast. [n]	3	20	17	16	3	8	4	3
Min maxkvot [$-$]	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Max maxkvot [$-$]	32,67	142,49	0,38	1,52	3,13	2,82	0,17	18,94
Median maxkvot [$-$]	0,01	0,21	0,09	0,02	0,19	0,11	0,02	0,09
Medel maxkvot [$-$]	2,03	0,64	0,13	0,24	0,25	0,15	0,03	0,11
STD_sam maxkvot [$-$]	7,88	6,28	0,11	0,42	0,27	0,12	0,03	0,29
STD_pop maxkvot [$-$]	7,76	6,27	0,11	0,40	0,27	0,12	0,03	0,29
Min minkvot [$-$]	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Max minkvot [$-$]	0,45	2,07	0,38	1,52	3,13	0,97	0,11	2,05
Median minkvot [$-$]	0,00	0,01	0,04	0,01	0,02	0,02	0,00	0,02
Medel minkvot [$-$]	0,05	0,13	0,10	0,17	0,11	0,05	0,01	0,04
STD_sam minkvot [$-$]	0,12	0,26	0,12	0,41	0,22	0,09	0,02	0,06
STD_pop minkvot [$-$]	0,11	0,26	0,12	0,40	0,22	0,09	0,02	0,06



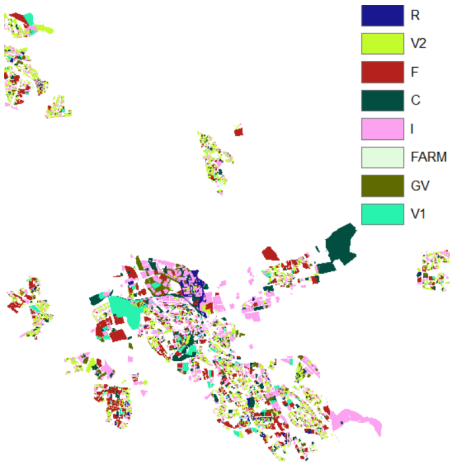
(a). 20%



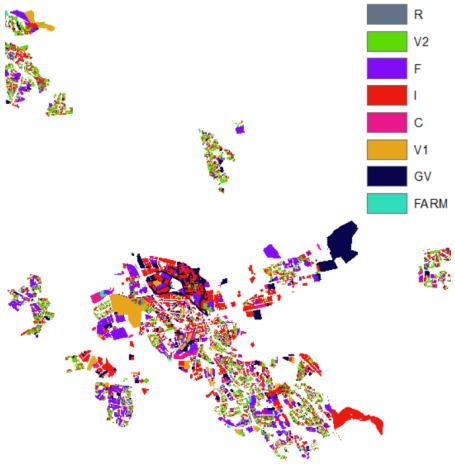
(b). 40%



(c). 60%



(d). 80%



(e). 100%

Figur C.1. *Färgkodade kartor över Linköping som visar klassificeringar baserade på största sannolikhet med 20%, 40%, 60%, 80% och 100% av tidigare klassificerade områden använda för att skapa signaturfiler.*

BILAGA D - NÄRMASTE GRANNE-ALGORITMEN OCH INTILLIGGANDE OMRÅDEN

Tabell D.1. *Data över närmaste granne-algoritmens prestation för olika bebyggelsetyper med tillgång till 5% redan klassificerat dataunderlag.*

Bufferradie ↓	Bebyggelsetyp →	C	F	FA	GV	I	R	V1	V2
50 m	Precision	0,07	0,97	0,00	0,07	1,00	0,93	0,00	0,96
	Känslighet	0,07	0,94	0,07	0,07	1,00	0,95	0,00	0,95
	F1	0,07	0,96	0,00	0,07	1,00	0,94	0,00	0,96
100 m	Precision	0,07	0,76	0,00	0,07	0,95	0,86	0,07	0,92
	Känslighet	0,00	0,78	0,00	0,00	0,85	0,87	0,00	0,91
	F1	0,00	0,77	0,00	0,00	0,90	0,86	0,00	0,92
200 m	Precision	0,07	0,66	0,00	0,07	0,86	0,78	0,25	0,83
	Känslighet	0,00	0,51	0,00	0,00	0,64	0,74	0,05	0,90
	F1	0,00	0,57	0,00	0,00	0,74	0,75	0,08	0,86
400 m	Precision	0,07	0,63	0,07	0,07	0,85	0,71	0,07	0,75
	Känslighet	0,00	0,39	0,00	0,00	0,53	0,63	0,00	0,88
	F1	0,00	0,48	0,00	0,00	0,65	0,67	0,00	0,81
600 m	Precision	0,07	0,64	0,07	0,07	0,87	0,69	0,07	0,72
	Känslighet	0,00	0,42	0,00	0,00	0,48	0,58	0,00	0,88
	F1	0,00	0,51	0,00	0,00	0,62	0,63	0,00	0,79
800 m	Precision	0,07	0,59	0,07	0,07	0,88	0,65	0,07	0,70
	Känslighet	0,00	0,38	0,00	0,00	0,43	0,58	0,00	0,85
	F1	0,00	0,46	0,00	0,00	0,58	0,61	0,00	0,77

Tabell D.2. *Data över närmaste granne-algoritmens prestation för olika bebyggelsetyper med tillgång till 10% redan klassificerat dataunderlag. Värden över 0,50 är markerade med en asterisk (*).*

Bufferradie ↓	Bebyggelsetyp →	C	F	FA	GV	I	R	V1	V2
50 m	Precision	0,07	0,96*	0,00	0,07	1,00*	0,94*	0,00	0,95*
	Känslighet	0,00	0,93*	0,07	0,07	1,00*	0,93*	0,00	0,96*
	F1	0,00	0,95*	0,00	0,07	1,00*	0,94*	0,00	0,96*
100 m	Precision	0,07	0,85*	0,00	0,07	0,86*	0,88*	0,00	0,91*
	Känslighet	0,00	0,80*	0,00	0,00	0,76*	0,85*	0,00	0,94*
	F1	0,00	0,82*	0,00	0,00	0,81*	0,87*	0,00	0,92*
200 m	Precision	0,07	0,71*	0,00	0,07	0,78*	0,81*	0,20	0,83*
	Känslighet	0,00	0,49	0,00	0,00	0,69*	0,74*	0,04	0,93*
	F1	0,00	0,58*	0,00	0,00	0,74*	0,78*	0,06	0,88*
400 m	Precision	0,07	0,63*	0,07	0,07	0,90*	0,75*	0,07	0,75*
	Känslighet	0,00	0,38	0,00	0,00	0,60*	0,61*	0,00	0,92*
	F1	0,00	0,48	0,00	0,00	0,72*	0,68*	0,00	0,83*
600 m	Precision	0,07	0,68*	0,07	0,07	0,92*	0,68*	0,07	0,72*
	Känslighet	0,00	0,40	0,00	0,00	0,51*	0,58*	0,00	0,88*
	F1	0,00	0,50	0,00	0,00	0,66*	0,63*	0,00	0,79*
800 m	Precision	0,07	0,64	0,00	0,07	0,89*	0,64*	0,07	0,70*
	Känslighet	0,00	0,29	0,00	0,00	0,43	0,59*	0,00	0,85*
	F1	0,00	0,40	0,00	0,00	0,58*	0,61*	0,00	0,77*

Tabell D.3. *Data över närmaste granne-algoritmens prestation för olika bebyggelsetyper med tillgång till 20% redan klassificerat dataunderlag. Värden över 0,50 är markerade med en asterisk (*).*

Bufferradie ↓	Bebyggelsetyp →	C	F	FA	GV	I	R	V1	V2
50 m	Precision	0,07	0,97*	0,07	0,00	0,93*	0,94*	0,07	0,95*
	Känslighet	0,07	0,86*	0,07	0,00	0,87*	0,93*	0,00	0,97*
	F1	0,07	0,91*	0,07	0,00	0,90*	0,94*	0,00	0,96*
100 m	Precision	0,20	0,81*	0,75*	0,00	0,85*	0,89*	0,50	0,91*
	Känslighet	0,10	0,74*	0,50	0,00	0,71*	0,85*	0,10	0,95*
	F1	0,13	0,77*	0,60*	0,00	0,78*	0,87*	0,17	0,93*
200 m	Precision	0,33	0,70*	0,71*	0,00	0,93*	0,83*	0,67*	0,85*
	Känslighet	0,06	0,56*	0,29	0,00	0,70*	0,77*	0,08	0,94*
	F1	0,11	0,63*	0,42	0,00	0,80*	0,80*	0,14	0,90*
400 m	Precision	0,07	0,64*	0,07	0,07	0,96*	0,60*	0,07	0,71*
	Känslighet	0,00	0,24	0,00	0,00	0,40	0,63*	0,00	0,81*
	F1	0,00	0,35	0,00	0,00	0,57*	0,62*	0,00	0,75*
600 m	Precision	0,07	0,61*	1,00*	0,07	0,95*	0,69*	0,07	0,74*
	Känslighet	0,00	0,42	0,05	0,00	0,57*	0,62*	0,00	0,87*
	F1	0,00	0,50	0,10	0,00	0,71*	0,65*	0,00	0,80*
800 m	Precision	0,07	0,65*	0,07	0,07	0,96*	0,64*	0,07	0,72*
	Känslighet	0,00	0,35	0,00	0,00	0,48	0,63*	0,00	0,83*
	F1	0,00	0,46	0,00	0,00	0,64*	0,63*	0,00	0,77*

Tabell D.4. *Data över närmaste granne-algoritmens prestation för olika bebyggelsetyper med tillgång till 40% redan klassificerat dataunderlag. Värden över 0,50 är markerade med en asterisk (*).*

Bufferradie ↓	Bebyggelsetyp →	C	F	FA	GV	I	R	V1	V2
50 m	Precision	0,00	0,96*	0,07	0,00	0,81*	0,94*	0,00	0,95*
	Känslighet	0,00	0,90*	0,07	0,00	0,81*	0,93*	0,00	0,97*
	F1	0,00	0,93*	0,07	0,00	0,81*	0,94*	0,00	0,96*
100 m	Precision	0,00	0,89*	1,00*	0,00	0,90*	0,90*	0,25	0,91*
	Känslighet	0,00	0,76*	0,50	0,00	0,76*	0,87*	0,07	0,95*
	F1	0,00	0,82*	0,67*	0,00	0,83*	0,88*	0,11	0,93*
200 m	Precision	0,25	0,75*	0,86*	0,07	0,94*	0,84*	0,50	0,85*
	Känslighet	0,07	0,61*	0,50	0,00	0,75*	0,78*	0,12	0,95*
	F1	0,11	0,67*	0,63*	0,00	0,84*	0,81*	0,19	0,90*
400 m	Precision	0,33	0,68*	1,00*	0,07	0,92*	0,79*	0,07	0,77*
	Känslighet	0,05	0,56*	0,08	0,00	0,67*	0,65*	0,00	0,93*
	F1	0,08	0,61*	0,14	0,00	0,77*	0,71*	0,00	0,84*
600 m	Precision	0,07	0,67*	0,07	0,07	0,94*	0,72*	0,07	0,72*
	Känslighet	0,00	0,49	0,00	0,00	0,59*	0,59*	0,00	0,89*
	F1	0,00	0,57*	0,00	0,00	0,72*	0,65*	0,00	0,80*
800 m	Känslighet	0,07	0,70*	0,07	0,07	0,94*	0,68*	0,07	0,72*
	Recall	0,00	0,38	0,00	0,00	0,49	0,62*	0,00	0,86*
	F1	0,00	0,49	0,00	0,00	0,64*	0,65*	0,00	0,78*

Tabell D.5. *Data över närmaste granne-algoritmens prestation för olika bebyggelsetyper med tillgång till 60% redan klassificerat dataunderlag. Värden över 0,50 är markerade med en asterisk (*).*

Bufferradie ↓	Bebyggelsetyp →	C	F	FA	GV	I	R	V1	V2
50 m	Precision	0,00	0,96*	0,07	0,07	0,94*	0,95*	0,00	0,95*
	Känslighet	0,00	0,91*	0,00	0,00	0,75*	0,93*	0,00	0,97*
	F1	0,00	0,93*	0,00	0,00	0,83*	0,94*	0,00	0,96*
100 m	Precision	0,00	0,86*	1,00*	0,07	0,95*	0,90*	0,50	0,91*
	Känslighet	0,00	0,75*	0,57*	0,00	0,74*	0,87*	0,17	0,96*
	F1	0,00	0,80*	0,73*	0,00	0,83*	0,89*	0,25	0,93*
200 m	Precision	0,75*	0,76*	1,00*	0,07	0,95*	0,83*	1,00*	0,85*
	Känslighet	0,30	0,60*	0,50	0,00	0,73*	0,78*	0,11	0,94*
	F1	0,43	0,67*	0,67*	0,00	0,82*	0,81*	0,19	0,89*
400 m	Precision	0,50	0,71*	1,00*	0,07	0,91*	0,80*	0,07	0,77*
	Känslighet	0,08	0,53*	0,09	0,00	0,65*	0,65*	0,00	0,94*
	F1	0,13	0,61*	0,17	0,00	0,76*	0,72*	0,00	0,85*
600 m	Precision	0,07	0,68*	0,07	0,07	0,93*	0,70*	0,07	0,72*
	Känslighet	0,00	0,47*	0,00	0,00	0,55*	0,59*	0,00	0,88*
	F1	0,00	0,55*	0,00	0,00	0,69*	0,64*	0,00	0,79*
800 m	Precision	0,07	0,72*	0,07	0,07	0,96*	0,67*	0,07	0,71*
	Känslighet	0,00	0,38	0,00	0,00	0,49	0,61*	0,00	0,86*
	F1	0,00	0,50	0,00	0,00	0,65*	0,64*	0,00	0,78*

Tabell D.6. *Data över närmaste granne-algoritmens prestation för olika bebyggelsetyper med tillgång till 80% redan klassificerat dataunderlag. Värden över 0,50 är markerade med en asterisk (*).*

Bufferradie ↓	Bebyggelsetyp ↓	C	F	FA	GV	I	R	V1	V2
50 m	Precision	0,00	1,00*	0,07	0,00	0,89*	0,95*	0,00	0,95*
	Känslighet	0,00	0,83*	0,00	0,00	0,73*	0,93*	0,07	0,97*
	F1	0,00	0,91*	0,00	0,00	0,80*	0,94*	0,00	0,96*
100 m	Precision	0,50	0,86*	0,07	0,07	0,95*	0,89*	1,00*	0,91*
	Känslighet	0,20	0,71*	0,00	0,00	0,64*	0,88*	0,14	0,96*
	F1	0,29	0,78*	0,00	0,00	0,77*	0,89*	0,25	0,94*
200 m	Precision	1,00*	0,79*	1,00*	0,07	0,94*	0,84*	0,07	0,84*
	Känslighet	0,33	0,62*	0,25	0,00	0,72*	0,77*	0,00	0,95*
	F1	0,50	0,70*	0,40	0,00	0,81*	0,81*	0,00	0,89*
400 m	Precision	0,50	0,69*	1,00*	0,07	0,92*	0,79*	0,07	0,77*
	Känslighet	0,14	0,47	0,25	0,00	0,63*	0,66*	0,00	0,94*
	F1	0,22	0,56*	0,40	0,00	0,75*	0,72*	0,00	0,85*
600 m	Precision	0,07	0,69*	0,07	0,07	0,90*	0,67*	0,07	0,72*
	Känslighet	0,00	0,42	0,00	0,00	0,52*	0,60*	0,00	0,87*
	F1	0,00	0,52	0,00	0,00	0,66*	0,63*	0,00	0,79*
800 m	Precision	0,07	0,75*	0,07	0,07	0,94*	0,64*	0,07	0,71*
	Känslighet	0,00	0,36	0,00	0,00	0,49*	0,60*	0,00	0,84*
	F1	0,00	0,48	0,00	0,00	0,65*	0,62*	0,00	0,77*

BILAGA E - EXPORT AV PRESTATIONSDATA

	Test Condition	Output Port
If	@Value(PREDICTED) = GR	@Value(PREDICTED) = GR
Else If	@Value(PREDICTED) = KOLONIOMRÅDE	@Value(PREDICTED) = KOLONIOMRÅDE
Else If	@Value(PREDICTED) = OBEBYGGD FASTIGHET	@Value(PREDICTED) = OBEBYGGD FASTIGHET
Else If	@Value(PREDICTED) = FOTBOLLPLAN	@Value(PREDICTED) = FOTBOLLPLAN
Else If	@Value(PREDICTED) = V1	@Value(PREDICTED) = V1
Else If	@Value(PREDICTED) = V2	@Value(PREDICTED) = V2
Else If	@Value(PREDICTED) = R	@Value(PREDICTED) = R
Else If	@Value(PREDICTED) = FARM	@Value(PREDICTED) = FARM
Else If	@Value(PREDICTED) = DAGBROTT	@Value(PREDICTED) = DAGBROTT
Else If	@Value(PREDICTED) = GV	@Value(PREDICTED) = GV
Else If	@Value(PREDICTED) = C	@Value(PREDICTED) = C
Else If	@Value(PREDICTED) = F	@Value(PREDICTED) = F
Else If	@Value(PREDICTED) = Ö	@Value(PREDICTED) = Ö
Else If	@Value(PREDICTED) = I	@Value(PREDICTED) = I
Else If	@Value(PREDICTED) = FLYGPLATS	@Value(PREDICTED) = FLYGPLATS
Else If		
Else	<All Other Conditions>	<UNFILTERED>

Figur E.1. *Skärmdump av testvillkor för korrekta klassificeringar i FME Desktop.*