Master Thesis
TVVR 20/5010

ANN-modell för att bestämma renoveringsbehov av vattenledningar

Utvärdering av viktiga attribut med tillämpning för Umeå kommun

Didrik Nilsson

Division of Water Resources Engineering
Department of Building and Environmental Technology
Lund University

ANN-modell för att bestämma renoveringsbehov av vattenledningar

Utvärdering av viktiga attribut med tillämpning för Umeå kommun

By:
Didrik Nilsson

Master Thesis

Division of Water Resources Engineering
Department of Building & Environmental
Technology

Lund University
Box 118
221 00 Lund, Sweden

Water Resources Engineering
TVVR-20/5010
ISSN 1101-9824

Lund 2020
www.tvrl.lth.se

Master Thesis
Division of Water Resources Engineering
Department of Building & Environmental Technology
Lund University

Swedish title:	ANN-modell för att bestämma renoveringsbehov av vattenledningar – Utvärdering av viktiga attribut med tillämpning för Umeå kommun
English title:	ANN-model to evaluate the need for maintenance of water pipes – Evaluation of important features based on the municipality of Umeå
Author:	Didrik Nilsson
Supervisors:	Johanna Sörensen, Erik Nilsson
Examiner:	Magnus Larson
Language:	Swedish
Year:	2020
Keywords:	underhåll, förnyelse, strategiskt, attributurval, maskininlärning

Förord

Jag vill rikta ett stort tack till de som hjälpt mig med mitt examensarbete: min handledare Johanna Sörensen på Lunds tekniska högskola var väldigt drivande i början av projektet och såg till att projektet hade en tydlig riktning och att rätt personer blev inblandade. Min handledare på Vakin, Petter Walan, var ett viktigt stöd genom hela arbetet där hans kunskap om Vakins databas och Umeås ledningsnät gjorde studien möjlig. Min biträdande handledare på Lunds tekniska högskola, Erik Nilsson, gjorde en utförlig granskning av rapporten och utan den hade rapporten både saknat viktigt innehåll och varit väldigt svår att tyda. Jag vill också tacka David Rehn på Stockholms vatten och avfall som först och främst bidrog med ett exempel på en ANN-modell för ett vattenledningsnät, men även bidrog med viktig återkoppling kring modellen som jag hade haft svårt att klara mig utan. Jag vill även tacka Vakin som bistått med allt jag behövt för att genomföra mitt examensarbete. Avslutningsvis vill jag tacka alla andra som hjälpt mig i mitt arbete.

Sammanfattning

Sveriges vattenledningsnät kräver kontinuerliga och stora investeringar och måste underhållas på ett effektivt sätt; syftet med den här studien var därför att utröna vilka ledningsattribut som är viktigast för att i en ANN-modell identifiera ledningar med hög risk för läckage. Detta gjordes genom att först använda attributurvalsmetoderna ReliefF och Recursive Feature Elimination (RFE) tillsammans med Random Forest Classification (RFC) och Multinomial logistisk regression (MLR) för att skapa urval av attributen. Dessa grupper av attribut användes sedan i ANN-modellen, och modellens prestation med respektive urvalsgrupp jämfördes. ReliefF och RFE med MLR lyckades rangordna attributen, medan RFE med RFC endast kunde särskilja tre attribut som mindre viktiga – dessa rangordningar innebar dock inte att attributen faktiskt var viktiga – de behövde testas i ANN-modellen först. Studien visade att antalet attribut kunde begränsas markant: när 10 utvalda attribut användes uppnåddes en noggrannhet på 0,79, att jämföra med alla tillgängliga attribut (19 stycken) då en noggranhet på 0,80 erhölls. Effekten av att endast inkludera attribut som är lätta att anskaffa och som är jämförbara mellan orter undersöktes också och en modellnoggrannhet på 0,75 uppnåddes då.

Abstract

Sweden's water pipe network demands continuous and large investments and must be maintained in an effective way; the aim with this study was therefore to investigate which pipe features are most important to identify pipes with a high risk of leakage in an ANN-model. This was done by first utilizing the feature selection methods ReliefF and Recursive Feature Elimination (RFE) with Random Forest Classification (RFC) and Multinomial Logistic Regression (MLR) to identify subsets of features seemingly important to evaluate the risk of leakage. The ANN-model were then run with these subsets and the difference in accuracy between subsets were compared. ReliefF and RFE with MLR succeeded in ranking features, while RFE with RFC only separated three features as less important—these rankings, though, had to be tested with the ANN-model to see if the features actually were important. The study found that the amount of features could be reduced distinctly: with 10 important features, an accuracy of 0.79 were achieved, to compare with 0.80 when all the 19 available features were utilized in the model. The effect of only including features that are easy to obtain and are similar between cities was also studied, and a model accuracy of 0.75 were obtained with these attributes.

Innehåll

Nomenklatur

Artificellt neuronnät (ANN) En maskininlärningsmetod som kan användas för klassificering.

Attribut Parametrar i ett prov. Attribut kan vara dimension, jordtyp, godstjocklek och så vidare.

Attributurval Processen att välja relevanta attribut. Processen benämns ”feature selection” på engelska.

Biasparameter Biasparametrar är parametrar utöver viktparametrar som adderas till data i varje lager.

Brus Data består av signal och brus. Brus är något som är irrelevant och minskar kvaliteten på data. I det här fallet kan brus vara attribut som inte bidrar med viktig information till modellen.

Filtermodell Attributsurvalsmetod där attribut väljs ut utan att ta hänsyn till den klassificeringsmodell som ska användas.

Inbäddad modell En inbäddad modell är en metodik för attributurval där filtermetodiken och wrappermetodikten kombineras. Frasen är översatt från engelskans ”embedded method”.

Manhattanavstånd Manhattanavstånd är avståndet som fås ifall avstånd i respektive dimension summeras. Noggrannare beskrivnings ges i avsnitt 2.3.1.

Multilayer Perceptron-modell (MLP-modell) En ANN-modell som består av minst tre lager.

Multinomial logistisk regression (MLR) En klassificeringsmetod som kan användas både för klassificering och för att undersöka korrelation. Kan användas med RFE för attributurval.

Prov I den här studien motsvarar ett prov en ledningssträcka. En samling prov används för att träna och utvärdera ANN-modellen.

Random Forest Classifier (RFC) En algoritm för klassificering som kan användas med RFE för attributurval.

Recursive Feature Elimination (RFE) En wrappermetod som tar bort ett anal attribut i varje körning.

Relief En filtermetod som används för attributurval.

ReliefF En variant på Relief som tar hänsyn till k närmsta grannar.

Viktparameter En ANN-modell består av lager där varje lager utgörs av viktparametrar. Träning av ett ANN innebär att dessa viktparametrar optimeras för att beskriva verkligheten. Noggrannare beskrivnings ges i avsnitt 2.2.

Wrapper Attributsurvalsmetod där attribut väljs ut baserat på hur klassificeringsmodellen presterar med och utan olika attribut. Jämför med filtermodeller där attributsurvalet görs utan att ta hänsyn till den klassificeringsmodell som ska användas.

1 Inledning

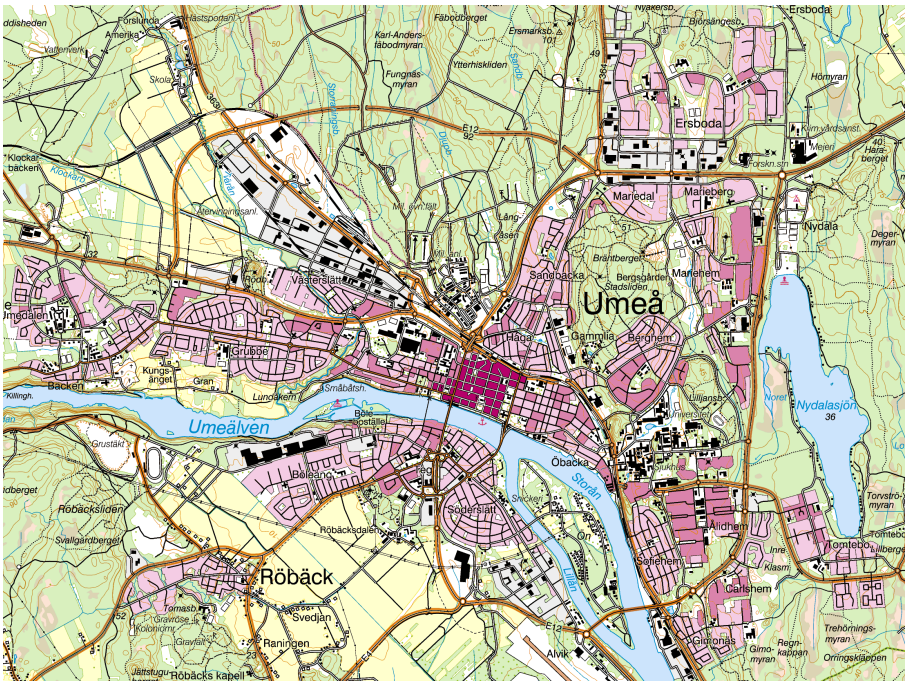
Distributionen av vatten är ett grundläggande samhällsbehov där stora investeringar gjorts och fortfarande görs. Återinvesteringskostnaden i det svenska ledningsnätet, vilket utöver dricksvattenledningar inkluderar spill- och dagvattenledningar, uppskattas till 600 miljarder kronor (Malm, Horstmark, Jansson m. fl., 2011), vilket nästan motsvarar två tredjedelar av svenska statens utgifter 2019 (Finansdepartementet, 2020). Ledningarna håller inte för evigt och måste underhållas, vilket med avseende på de stora kostnaderna måste ske på ett strategiskt sätt. Eftersom vattenledningssystem är förlagda i mark är det svårt och dyrt att undersöka ledningarnas kondition, även om det i viss mån är möjligt. Därför är det av intresse att identifiera alternativa metoder som kan nyttja mer lättåtkomlig information. Ett angreppssätt är att nyttja en modell baserad på ett artificiellt neuronnät (ANN), vilket undersöks i forskningsprojektet *Ordning i RörANN*. Projektet är ett samarbete mellan Lunds tekniska högskola, Sweden Water Research och de kommunala bolagen Stocholm vatten och avfall (SVOA), VA syd, Vakin (Umeå) och Kretslopp och vatten (Göteborg) där projektets mål är att utveckla en ANN-modell som går att applicera på Sveriges kommuner. Projektet bygger på en modell utvecklad för Stockholm av SVOA där underhållsbehovet uppskattas genom 21 inparametrar (Rehn & Giertz, 2019). Många kommuner har dock inte tillgång till alla dessa 21 inparametrar, vilket gör användningsområdet för modellen begränsat. Syftet med det här examensarbetet är därför att försöka begränsa antalet inparametrar som behövs för att uppnå ett acceptabelt resultat. Mer konkret ämnar arbetet att svara på de tre frågorna:

- Vilka metoder kan användas för att utvärdera viktiga attribut i en ANN-modell för att beskriva ett dricksvattennäts kondition?
- Vilka attribut är viktigast för att beskriva ett dricksvattennäts kondition?
- Är det möjligt att med attribut tillgängliga även för kommuner med lite data erhålla en modell som presterar i nivå med en modell

tränad på attribut tillgängliga för en kommun med god tillgång till data?

1.1 Fallstudie

Umeå kommun ligger i Västerbotten och har en befolkningsmänd på 129 000 invånare[1], med en tätort på 114 000 invånare [2] . Umeå ligger vid kusten och har en normal årstemperatur på 2 grader Celsius. En karta över Umeå stad visas i figur 1



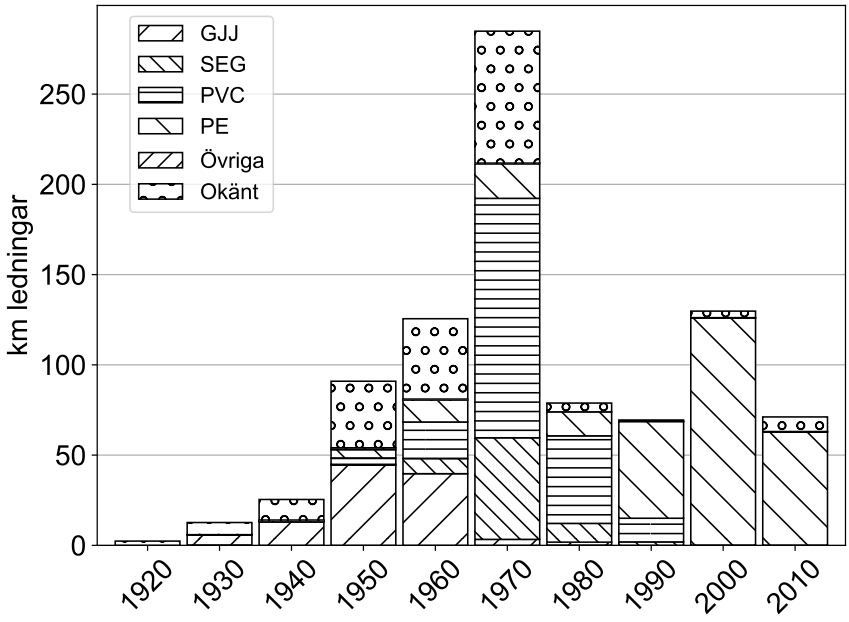
Figur 1. En terrängkarta över Umeå stad (Geografiska Sverigedata, 2016).

Umeå har 900 km huvud- och distributionsledningar och dess ålders- och materialfördelning redovisas i figur 2. De senaste 10 åren har Umeå haft ett genomsnittligt utläckage av dricksvatten på 20 procent (Walan,

[1]https://www.umea.se/umeakommun/kommunochpolitik/kommunfakta.4.bbd1b101a585d704800061691.html

[2]http://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__MI__MI0810__MI0810A/BefLandInvKvmTO/table/tableViewLayout1/

2019). 14 procent av dessa är utgörs av omätt vatten. Detta innebär att Umeå ligger på en nationell normalnivå vad gäller utläckage.



Figur 2. Antal kilometer lagda per decennium i Umeå.

Antalet läckor per år varierar mycket beroende på vädret. Hur antalet läckor fördelat sig över de senaste tio åren redovisas i tabell 1. Dessa siffror är baserade på den utsökning som gjorts i den här studien och inte på officiell statistik från VASS. Detta för att följa definitionen av läcka som använts i den här rapporten. Främst åren 2010 och 2011 avviker från VASS-statistiken. Inrapporteringen av dessa läckor skilde sig från resterande år och därför missades de läckorna i den här studien.

Tabell 1. Antal vattenläckor per år i Umeå kommun baserat på den utsökning som gjordes i den här studien.

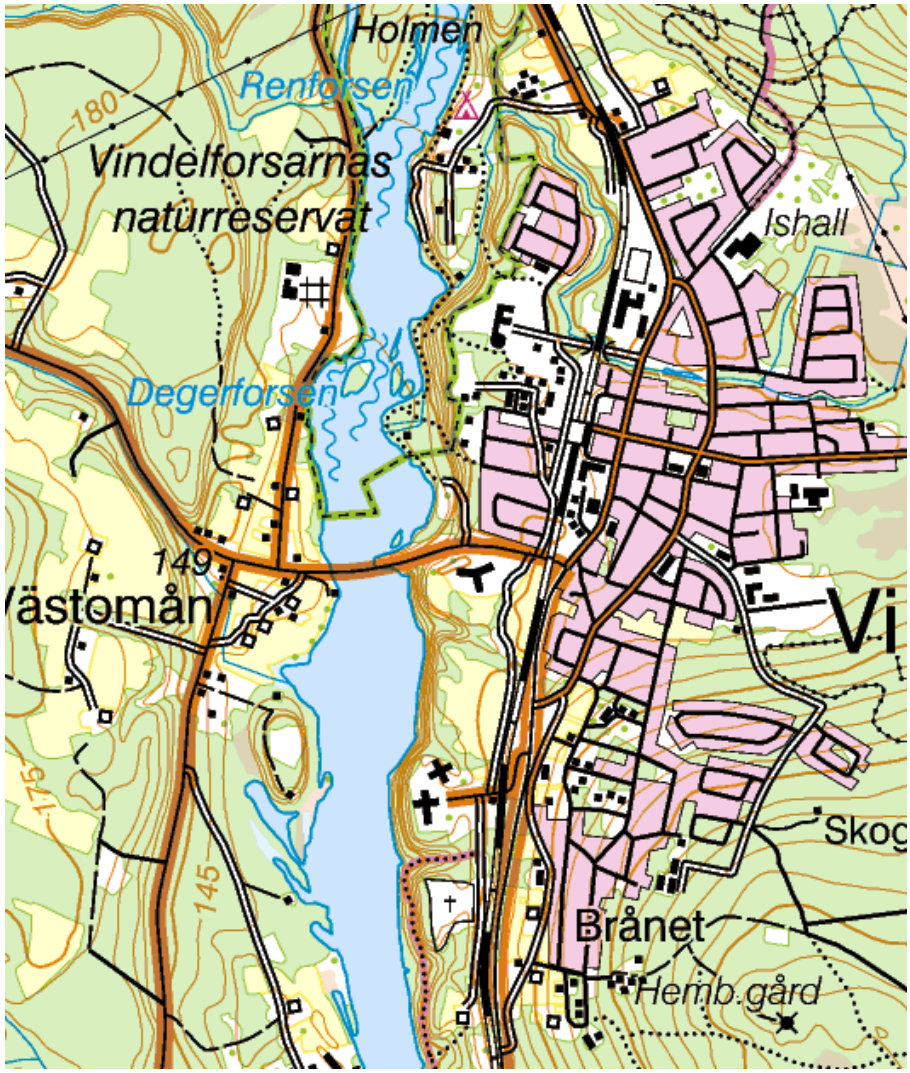
År	Antal läckor
2010	5
2011	8
2012	39
2013	67
2014	73
2015	76
2016	119
2017	72
2018	110
2019	93

Umeå har för stunden ingen underhållsplan men det är något som Vakin håller på att ta fram. Det finns dock en förnyelseplan där strategier för förnyelse av ledningsnätet presenteras. I förnyelseplanen fastslås att Umeås nuvarande förnyelsetakt är för låg och måste öka från 0,35 procent (baserat på åren 2012 – 2016) till ca 0,8 procent per år (Walan, 2019). Vilka delar av ledningsnätets som prioriteras bygger på två modeller från Svenskt Vatten: dels en förnyelsetakt för respektive material, dels en områdesvis förnyelsetakt. Den områdesvisa modellen har justerats och överensstämmer inte helt med Svenskt Vattens grundmodell[3]. Dessutom har åtgärder vidtagits för att kunna arbeta proaktivt med läcksökning där både vattenmätare som lyssnar efter läckor och flödesmätare på utvalda platser installerats (Walan, 2019). Mycket förnyelse har historiskt baserats på samverkan med andra aktörer där ledningar exempelvis bytts ut i samband med vägomläggningar.

Vindelns kommun är en liten kommun sex mil in i landet från Umeå med 5400 invånare[4] och har en normal årstemperatur på 1 grad Celsius. En karta över Vindelns samhälle visas i figur 3.

[3]Petter Walan, Utredningsingenjör på Vakin

[4]https://www.vindeln.se/Sve/Filarkiv/Kommunfakta%20och%20Organisation/Kommunfakta%202017.pdf



Figur 3. En terrängkarta över Vindelns samhälle (Geografiska Sverigedata, 2016).

Datatillgängligheten i Vindeln är begränsad jämfört med Umeå och många ledningsattribut är okända. Denna kommun kommer att användas för att undersöka möjligheten att använda en ANN-modell när datatillgången är sämre. Modellen kommer ej att köras på Vindelns kommun, utan en uppskattning av tillgängliga attribut för Vindelns kommun kommer att göras och modellen utvärderas med dessa fast med Umeås data. Vindelns kommun har i dag ingen förnyelseplan utöver en

prioriteringslista över ledningar. Vakin håller dock på att utveckla en strategi för Vindeln under rapportens skrivande.

1.2 Avgränsningar

Läcka kommer att ha en vital roll i det här arbetet och därför är det viktig att definiera vad som menas med en läcka. Inrapporteringsprogrammet VASS använder följande definition för att beskriva en vattenläcka för inrapportering (VASS, 2020):

> Antal reparerade rörbrott/vattenläckor på det allmänna vattenledningssystemet under året, exkl serviser. Även läckor på ventiler och brandposter ingår, liksom läckor som hittats vid aktiv läcksökning.

Eftersom läckor rapporteras in till VASS på det sättet är det rimligt att utgå ifrån den definitionen för läcka, men eftersom det kan antas vara andra orsaker till läckor på ventiler och brandposter än ledningar, inkluderas inte dessa läckor i begreppet *läcka* i den här rapporten.

2 Teori

2.1 Vattenledningar

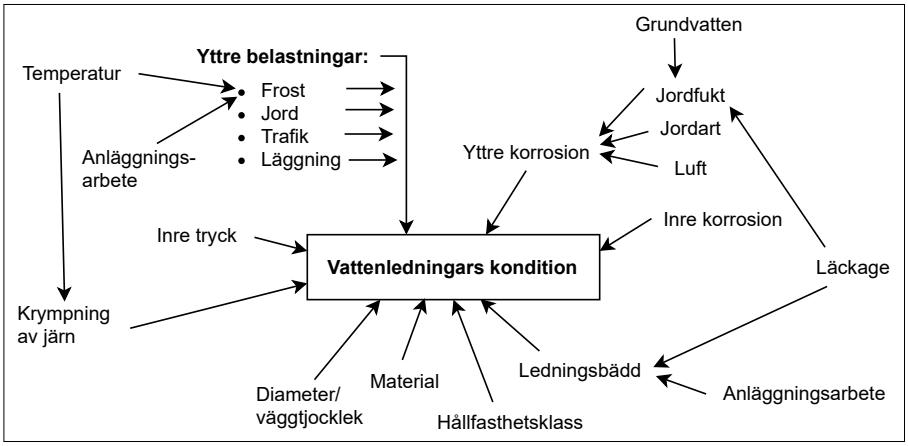
För att definiera en läcka i den här rapporten användes VASS definition (VASS, 2020). För att utvärdera ett ledningsnät vore det dock fördelaktigt ifall modellen även identifierade ledningar med dålig kondition. Att fokus läggs på läckage beror på att det är ett tydligt mått på att en ledning är i behov av underhåll – eftersom vattenledningar är trycksatta är det svårt att mäta ledningens status på något annat sätt. Läckor kan vara ett resultat av dålig kondition på en ledning, och därför är det intressant att förstå faktorer som påverkar konditionen på vattenledningar, och delavsnitt 2.1.1 ger därför en kortfattad genomgång av vad som påverkar en lednings status och livslängd. Resterande del av avsnittet behandlar förhållningssätt till underhåll och strategier som kan användas.

2.1.1 Ledningars kondition

En vattenlednings funktion bedöms utifrån två kriterier: vattenkvalitet och ledningsfunktion (Lidström, 2013; Malm, Horstmark, Jansson m. fl., 2011). Det finns flera olika faktorer som påverkar ledningars funktion och Rajani och Kleiner (2001) delar upp dessa faktorer i tre grupper: förutsättningar, laster och nedbrytning. I dessa grupper ingår:

- Förutsättningar:
  - ledningens struktur och material
  - samverkan mellan ledning och fyllnadsmaterial
  - kvalitet på läggningsarbete.
- Laster:
  - interna
  - externa (jordlast, trafiklast, tjäle)
  - specialfall såsom avgrävning.
- Nedbrytning:
  - kemisk
  - biologisk
  - elektrokemisk.

Detta stämmer väl överens med faktorer indikerade av Clark m. fl. (1987), men där nämns även explicit väder och anslutningar. Figur 4 illustrerar olika påverkansfaktorer på ledningar, och i tabell 2 visas en sammanställning av orsaker till läckor för ledningar i Sverige (Sundahl, 1996). Trots figurens ålder är den fortfarande aktuell – PE är det vanligaste materialet i dag, och användes även 1996. Dessutom är många ledningar i ledningsnätet äldre än 25 år. Flera av dessa faktorer är dock svåra att inkludera i en modell då exempelvis *kvalitet på läggningsarbete* inte går att mäta i efterhand.



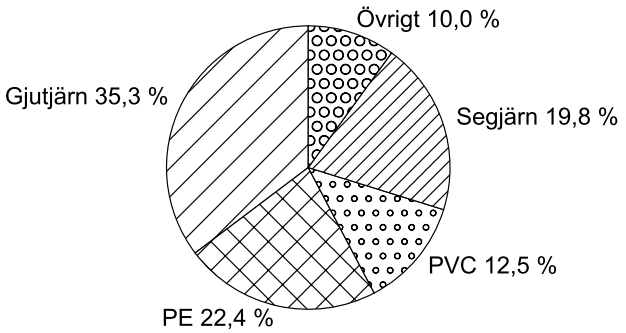
Figur 4. Olika faktorer som påverkar konditionen på en vattenledning. Omarbetat från Sundahl (1996).

Tabell 2. Feltyper och orsaker för läckage för ledningsmaterialen gjutjärn, segjärn, PVC och PE. Värdena är baserade på data från Eskilstuna, Luleå, Malmö, Västerås och Örebro. Tidsperioden varierar mellan stad och material, men totalt finns data från 1965 till 1993. Omarbetat från Sundahl (1996).

Feltyp och orsak	Gjutjärn %	Segjärn %	PVC %	PE %
Sprickor eller brott med orsak okänd	55	50	18	10
Sprickor eller brott med orsak sättning	24	14	4	5
Korrosion	11	16	0	0
Fogfel	2	5	76	48
Tjäle eller frysning	2	2	0	0
Okänt	6	13	2	37

Ledningsmaterialet har en stor inverkan både på risken för brott och typen av brott. I Sverige är det främst fyra ledningsmaterial som dominerar i det existerande ledningsnätet: PE, PVC, segjärn och gjutjärn (Malm & Svensson, 2011) vilket illustreras i figur 5. Vissa faktorer påverkar ledningskonditionen oavsett material där ledningar med små diametrar är mer utsatta för brott än ledningar med stor diameter (Sundahl, 1996). Vidare anges även årstid vara en viktigt faktor för skador på

ledningar, där främst den lägre temperaturen på vintern bidrar till fler läckor. Utöver detta kan även rötter påverka ledningar.



Figur 5. Ledningsmaterial i existerande ledningsnät 2009. Omarbetat från Malm och Svensson (2011)

Ledningar i metall kommer med tiden att korrodera både invändigt och utvändigt, där yttre korrosion generellt leder till fler läckor än inre (Lidström, 2013). Korrosion uppstår då elektroner kan röra sig från en anodyta till en katodyta (Stål & Wedel, 1984), och korrosionshastigheten varierar med den miljö en ledning placeras i – vid exempelvis hög grundvattenyta, eller jord som håller kvar vatten länge, kommer korrosionshastigheten att öka (Sægrov, 1998). Ifall en lednings korrosionsskydd lokalt försvinner agerar en stor del av ledningen katod för den lilla anod som uppstått, och det är främst denna lokala korrosion som är viktig vid brott (Sægrov, 1998).

För gjutjärn har många brott kopplats till trafiklaster, jordlaster samt balkverkan från undermålig ledningsbädd (Sægrov, 1998), och i Malm, Horstmark, Larsson m. fl. (2011) anges ålder och sättningar som stora orsaker till brott för gjutjärn.

I Sægrov (1998) anges korrosion vara den största anledningen till brott för segjärn. En anledning till att korrosion upplevts vara ett större problem för segjärn än gjutjärn kan vara på grund av att segjärn på grund av bättre hållbarhet kan läggas med tunnare godstjocklek och därför har en mindre buffert mot korrosion (Malm, Horstmark, Larsson m. fl., 2011).

De två vanliga plastmaterialen, PE och PVC, är båda termoplaster.

Termoplaster har en tendens att krypa vilket för ledningar leder till spänningar i ledningen (Malm, Horstmark, Larsson m. fl., 2011), där spänningar påverkar risken för brott. Plasternas brottyper delas in i tre grupper: (1) när ledningen är ung uppstår segt brott för att sedan (2) gå över till sprött brott när ledningen blir äldre. Den maximala livslängden begränsas av det tredje brottillståndet (3): åldersprickor. Detta sker dock efter flera hundra år vid normala förhållanden. För plastledningars livslängd är i teorin framförallt de fyra faktorerna materialets egenskaper, belastning, temperatur och miljö viktiga. I praktiken är det dock endast belastning och materialets egenskaper som är relevanta, där godstjockleken är speciellt viktig (Malm, Horstmark, Larsson m. fl., 2011).

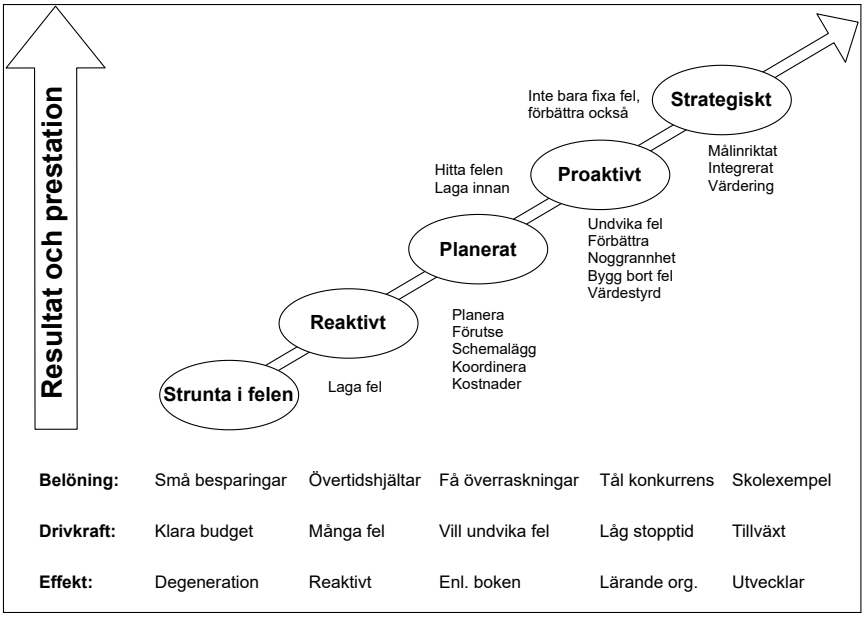
Gamla PVC-ledningar har varit starkt förknippade med läckor. Först var materialet undermåligt (Malm, Horstmark, Larsson m. fl., 2011) men då detta korrigerats började i stället ledningarna sammanfogas med en speciell muff, Ehri-muffen, som inte fungerade tillfredsställande (Sundahl, 1996). De PVC-ledningar som lagts efter perioden med Ehri-muff presterar väl (Malm, Horstmark, Larsson m. fl., 2011). PVC-ledningar bör i teorin ge sega brott men i praktiken observeras i stället spröda brott (Sægrov, 1998). Detta har hänförts till yttre faktorer såsom trafiklast och dåligt utförd ledningsbädd.

Hållbarheten för plasten PE påverkas av temperatur, syrehalt, fukthalt och mikroorganismer (Hakkarainen & Albertsson, 2004). Detta påverkar den totala livslängden för plasten, men samtliga faktorer är inte relevant för den miljö som vattenledningar befinner sig i. Nedbrytningen på grund av syre uppmärksammas dock i Sægrov (1998), men denna nedbrytning begränsas med antioxidanter i materialet (Hakkarainen & Albertsson, 2004; Sægrov, 1998). De faktiska brotten i ledningar av PE uppstår främst från mikrosprickor som växer sig större (Sægrov, 1998). I Sverige är den främsta anledningen för läckage fogskador, där övergång från större till mindre ledning är mest kritisk (Malm, Horstmark, Larsson m. fl., 2011). Vilken svetsfog som används påverkar hållfastheten.

2.1.2 Underhållsstrategier

Återinvesteringskostnaden för det svenska VA-systemet uppskattas till 800 miljarder kronor, varav 600 miljarder kronor kopplas till ledningsnätet (Malm, Horstmark, Jansson m. fl., 2011). Ledningarna har en begränsad livslängd vilket innebär att en stor summa pengar kontinuerligt måste investeras i Sveriges ledningsnät, och för att detta ska kunna ske effektivt måste det finnas underlag för vilka ledningar som ska bytas ut eller förbättras. Om förnyelsetakten är för låg kommer kostnaden för ledningsnätet bli större i framtiden. Ett attribut som har inverkan är ledningens ålder, och exempelvis en teknisk medellivslängd på 80 år skulle innebära att ledningsförnyelsen skulle behöva vara 1/80=1,25 procent per år, att jämföra med dricksvattennätets medianförnyelsehastighet i Sverige på 0,4 procent (Malm, Horstmark, Jansson m. fl., 2011). Av olika anledningar, bland annat fördelningen mellan gamla och nya ledningar, skulle en förnyelsetakt på 1,25 procent innebära att ledningarna aldrig uppnår den tekniska livslängden med en för hög investeringstakt med onödiga kostnader som resultat. I stället är det relationen mellan ledningars status och mål för ledningsnätet som bör ligga till grund för underhåll (Malm, Horstmark, Jansson m. fl., 2011).

Det finns flera principiellt olika förhållningssätt till drift och underhåll av ledningsnät. Förenklat kan underhållsstrategierna delas in i fem grupper: *strunta i felen*, *reaktivt*, *planerat*, *proaktivt* och *strategiskt* underhåll, där *strunta i felen* är sämst för ledningsnätet och *strategiskt underhåll* bäst (Jacobsson m. fl., 2019). Orsaker till varför en ledningsägare hamnar i en specifik grupp kan vara flera, men förenklat kan drivkraften för *strunta i felen* vara att klara budget medan för *strategiskt underhåll* i stället vara tillväxt. Detta illustrerats i figur 6.



Figur 6. Olika sorters förhållningssätt till underhåll på vattenledningar, samt deras implikationer. Omarbetat från Jacobsson m. fl. (2019).

För att arbeta systematiskt med ett ledningsnät krävs någon form av strategi och i Malm, Horstmark, Jansson m. fl. (2011) anges fyra metoder för systematisk underhållsplanering (de fyra metoderna är på en översiktlig nivå). Den första metoden är att basera förnyelsetakten på uppskattad teknisk livslängd, vilket som beskrivits tidigare är vanskligt. Den andra metodiken är en enkel metod där ledningsnätet förnyas med en konstant hastighet varje år, där den lägsta förnyelsetakten ej bör understiga 0,3 procent (Stahre m. fl., 2007). Efter fem år utvärderas ledningsnätets skick och ifall driftstörningarna börjat öka, ökas förnyelsetakten med 0,2 procentenheter – exempelvis från 0,3 till 0,5 procent. Om driftstörningarna fortsätter att öka, ökas förändringstakten återigen med 0,2 procentenheter, och så vidare.

I den tredje metoden delas ledningsnätet in i gamla och nya ledningar (Malm, Horstmark, Jansson m. fl., 2011). För de gamla ledningarna undersöks driftstörningsfrekvensen med avseende på ifall denna är relativt konstant eller förändras, och ifall nivån är acceptabel. Om mängden driftstörningar är acceptabel och konstant vidtas ingen åtgärd.

Om nivån inte är acceptabel, bestäms andelen *för många driftstörningar*, vilket ligger till grund för beslut om underhåll. För de nya ledningarna bedöms förnyelsetakten utifrån ålder och hur mycket som måste förnyas; ofta ökar denna andel med tiden och ett sätt att uppskatta förnyelsetakten redovisas i ekvation 1 där N är antalet år efter att ledningen lagts.

$$\text{Förnyelsetakt (efter N år)} = \frac{\text{N}}{(\text{Medianlivslängd})^2} \tag{1}$$

Denna empiriska formel gäller upp till medianlivslängd och efteråt måste någon annan bedömning göras, förslagsvis baserat på antalet driftstörningar (Malm, Horstmark, Jansson m. fl., 2011).

Den fjärde och sista metoden utnyttjar ledningars ålder och hur länge de bör hålla för att beräkna förnyelsetakten (Malm, Horstmark, Jansson m. fl., 2011). Svenskt Vatten tillhandahåller en Excelfil för att underlätta denna utvärdering, men andra program finns också tillgängliga.

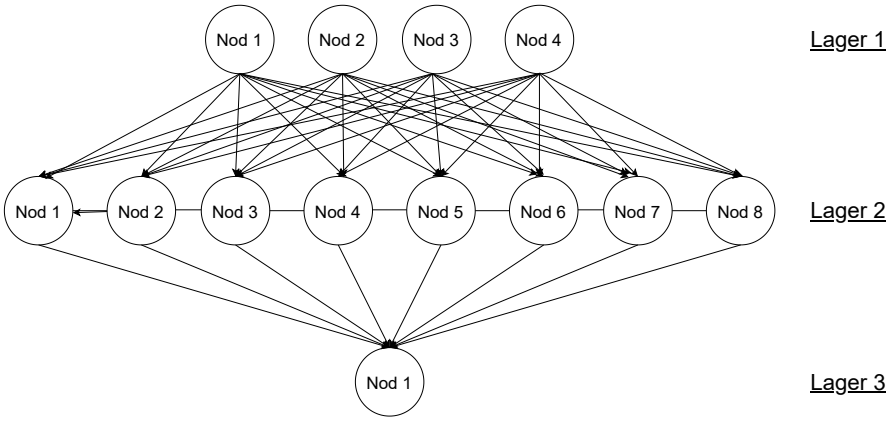
2.2 Artificiella neuronnät

Tre av de fyra tidigare nämnda metoderna för att bedöma konditionen på ett ledningsnät är enkla tumregler. För att öka beslutsunderlaget finns det behov av ytterligare utvärderingsmetoder; en sådan kan vara artificiella neuronnät, ANN, som har använts av SVOA i Stockholm (Rehn & Giertz, 2019). En fördel med en ANN-modell är att den indikerar underhåll på objektsnivå till skillnad från de andra utvärderingsmetoderna. ANN-modellen tränas på delar av det befintliga ledningsnätet för att kunna identifiera ledningar med läckage. För att träna modellen används en kombination av ledningar med och utan observerad läcka. Även om modellen tränas på observerade läckor då det är ett objektivt mått, är förhoppningen att modellen ska kunna identifiera ledningar med små oidentifierade läckor eller där ledningsnätet har dålig kondition. Detta för att ANN-modellen ska kunna användas som hjälpmedel vid strategiskt underhåll. Då modellens prestation fortsättningsvis diskuteras kommer endast identifierade och oidentifierade läckor att användas i och med måttet läckas objektivitet. Hur en ANN-modell fungerar och vad det innebär att modellen tränas beskrivs i det här avsnittet.

ANN används inom flera områden för att identifiera samband mellan attribut och utfall vid stora datamängder. Ett vanligt användningsområde för ANN-modeller är bildigenkänning – en modell kan exempelvis läras att urskilja ifall en djurbild är en hund eller inte. Detta fungerar genom att modellen får se väldigt många bilder på hundar och inte hundar. För varje bild är det också angivet ifall bilden är en hund eller inte. Modellen får sedan se dessa bilder om och om igen för att modellen ska lära sig att känna igen karaktäristika för en hund. Detta är vad som åsyftas med att träna en modell. När modellen tränas på alla dessa bilder där djuret är känt, kan modellen sedan användas för att klassa bilder som den inte sett tidigare. Det modellen analyserar är egentligen inte bilden, utan bildens pixelvärden (varje pixels färg representeras av en siffra). En ANN-modell för ett dricksvattennät fungerar på exakt samma sätt, men i stället för pixelvärden analyserar modellen ledningsattribut såsom diameter och jordmaterial. Den viktigaste skillnaden mellan de två modellerna är att när modellen tränas på hundbilder kan väldigt mycket träningsdata genereras. För ett ledningsnät är mängden data i stället begränsad vilket försvårar inlärningen. Dessutom är det omöjligt att ta hänsyn till alla faktorer som påverkar ett ledningsnät – det är inte lika entydigt som en grupp pixelvärden.

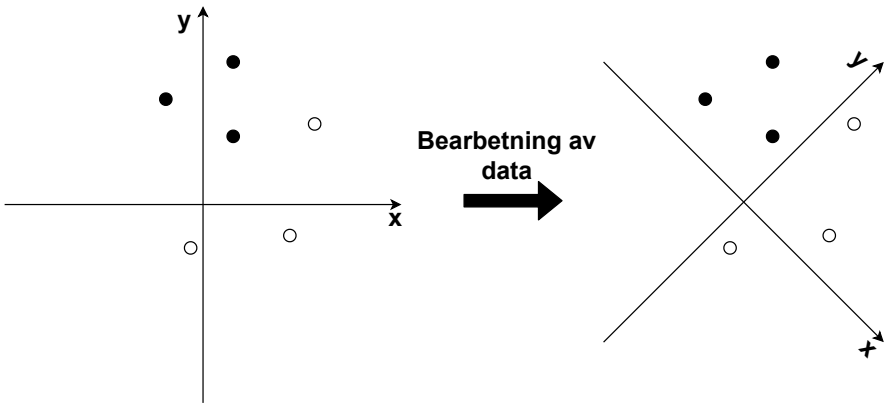
Ett vanligt förklaringsförsök är att likna processen för ett ANN med hur information överförs och behandlas i hjärnan, men detta är inkorrekt och kan leda till förvirring, och i det här arbetet används en mer matematisk beskrivning. En ANN-modell består av ett antal lager och inom varje lager finns det noder där någon form av omvandling av data sker (Chollet, 2017). Figur 7 illustrerar en enkel ANN-struktur som består av tre lager, varav det första lagret består av fyra noder, det andra lagret av åtta noder och det sista lagret av en nod. Antalet noder i det första lagret bestäms av antalet attribut i indata, och antalet noder i det sista lagret bestäms av hur många dimensioner utdata ska vara i; i det här fallet har indata fyra dimensioner och utdata en dimension. Om detta hade varit en ANN-modell för ett ledningsnät, hade indata varit fyra ledningsattribut, exempelvis ledningsdjup, längd, ålder och diameter, och utdata hade varit ett värde som anger om ledningen är i behov av underhåll eller inte. Antalet noder i resterande lager kan väljas fritt, men

för många noder kan leda till att modellen lär sig samband som endast gäller för det aktuella datasetet (överpassning), medan för få noder inte tillåter modellen att finna tillräckligt komplexa samband (Chollet, 2017).



Figur 7. En enkel illustration av ett artificiellt neuronnät.

Omvandlingen av data i respektive lager kan vara av flera typer, exempelvis transformationer och rotationer. Figur 8 visar hur en rotation kan användas för att klassificera data: efter rotationen kan vita prickar väljas genom att ta alla prickar där x är större än noll. Utöver omvandling används en transferfunktion som ökar antalet möjliga utfall – utan transferfunktioner skulle endast linjära operationer kunna genomföras.



Figur 8. Genom att rotera koordinatsystemet kan vita prickar enkelt särskiljas från svarta. Omarbetat från Chollet (2017).

Matematiskt kan omvandlingen i ett lager beskrivas som i ekvation 2 där *transf* är en en transferfunktion, X_{k-1} är utdata från det föregående lagret, V_k är en viktmatris i det aktuella lagret, B_k är en bias-vektor i det aktuella lagret och X_k är de omvandlade värdena från det aktuella lagret (Keras, 2019). Biasvektorn ökar modellens inlärningsmöjligheter eftersom den möjliggör addition utöver multiplikation.

$$X_k = transf(X_{k-1} \times V_k + B_k) \tag{2}$$

Som exempel kan strukturen i figur 7 återigen användas. Indata till lager två har fyra attribut och utdata från lager två har åtta attribut. X_{k-1} är då en vektor med dimensionen 1×4, V_k är en matris med dimensionen 4×8, B_k är en biasfaktor med dimensionen 1×8 och X_k är en vektor med dimensionen 1×8.

Det finns flera olika transferfunktioner, och en enkel variant sätter alla värden över ett tröskelvärde till ett och alla värden under tröskelvärdet till noll (Chollet, 2017). Vilken transferfunktion som bör användas varierar från fall till fall och från lager till lager, men för att bestämma ifall en ledning behöver underhåll eller inte bör den sista transferfunktionen vara en funktion som möjliggör binär klassning, exempelvis en tröskelfunktion. I Stockholmsmodellen (Rehn & Giertz, 2019), som är den modell som detta arbetet baseras på, används en sigmoid funktion, och en sigmoid funktion är en funktion, $f(x)$, som går från nära noll till nära ett över ett kort spann på x-axeln (Chollet, 2017). Den sigmoida funktionen omvandlar x-värdet till ett värde mellan noll och ett enligt $X_n = transf(X_{n-1} \times V_n + B_n)$ (där n är det totala antalet lager). Den sigmoida funktionen $transf(x)$ är kontinuerlig, men $transf(x) \geq 0,5$ klassas till 1 och $f(x) < 0,5$ klassas till 0 (då utfallet jämförs med facit). Den sigmoida funktionen ger kontinuerliga värden mellan 0 och 1, och kan därför utöver att användas för klassificering, tolkas som en form av sannolikhet för att ett prov tillhör en klass. Detta är användbart när en modell ska utvärderas, där ett värde nära 1 sannolikt är 1, medan ett värde precis över 0,5 kan tolkas vara mer osäkert om det ska tillhöra klassningen 1 eller 0 (Chollet, 2017).

För att koppla detta till noderna kan figur 7 användas igen. Som kan

ses går det pilar från varje nod i ett lager till alla noder i nästkommande lager (att alla noder mellan två lager är kopplade till varandra kallas att lagret är *fullt anslutet*, men det är inget krav) (Chollet, 2017). Detta innebär att vektorn X_{k-1} går till alla noder. I varje nod sker sedan en matrismultiplikation enligt ekvation 2. För nod två i lager två enligt figur 7 skulle denna operation se ut som i ekvation 3. Utdata från lager två blir en vektor med samma längd som antalet noder i lager två, där första termen i vektorn är resultatet av operationen i nod ett, andra termen av operationen i nod två och så vidare.

$$
\begin{aligned}
&transf\left(\begin{bmatrix} x(1)_1 & x(1)_2 & x(1)_3 & x(1)_4 \end{bmatrix} \times \begin{bmatrix} v(2)_{1,2} \\ v(2)_{2,2} \\ v(2)_{3,2} \\ v(2)_{4,2} \end{bmatrix} + \begin{bmatrix} b(2)_2 \end{bmatrix} \right) \\
&= \\
&transf\left(\begin{bmatrix} \big(x(1)_1 \times v(2)_{1,2}\big) + \big(x(1)_2 \times v(2)_{2,2}\big) + \\ +\big(x(1)_3 \times v(2)_{3,2}\big) + \big(x(1)_4 \times v(2)_{4,2}\big) + b(2)_2 \end{bmatrix} \right) \\
&= \\
&\begin{bmatrix} x(2)_2 \end{bmatrix}
\end{aligned}
\tag{3}
$$

där

$x(1)_i$ är värdet på utdata från nod i i lager 1

$v(2)_{i,2}$ är viktparameter i i nod 2 för lager 2

$b(2)_2$ är bias-vektorn för nod 2 i lager 2

$x(2)_2$ är värdet på utdata från nod 2 i lager 2.

Varje nods viktvektor är en kolumn i lagrets viktmatris. Relationen mellan viktvektor och viktmatris illustreras i ekvation 4 för det andra

lagret i figur 7.

$$\text{Viktvektor nod 2} = \begin{bmatrix} \mathbf{v_{1,2}} \\ \mathbf{v_{2,2}} \\ \mathbf{v_{3,2}} \\ \mathbf{v_{4,2}} \end{bmatrix}$$

$$\text{Position i lagrets viktmatris} = \tag{4}$$

$$\begin{bmatrix} v_{1,1} & \mathbf{v_{1,2}} & v_{1,3} & v_{1,4} & v_{1,5} & v_{1,6} & v_{1,7} & v_{1,8} \\ v_{2,1} & \mathbf{v_{2,2}} & v_{2,3} & v_{2,4} & v_{2,5} & v_{2,6} & v_{2,7} & v_{2,8} \\ v_{3,1} & \mathbf{v_{3,2}} & v_{3,3} & v_{3,4} & v_{3,5} & v_{3,6} & v_{3,7} & v_{3,8} \\ v_{4,1} & \mathbf{v_{4,2}} & v_{4,3} & v_{4,4} & v_{4,5} & v_{4,6} & v_{4,7} & v_{4,8} \end{bmatrix}$$

Detta är modellens strukturella, förenklade matematiska uppbyggnad. För att omvandlingen av indata ska kunna ge meningsfull utdata måste modellen tränas, vilket kräver indata där det även finns ett facit – det vill säga vad utdata ska bli givet indata. Detta är en iterativ process där de initiala viktparametrarna och biasvektorerna först slumpas fram för att sedan i varje iteration ändras för att omvandlingen av indata bättre ska representera utdata. En modells förlust är ett mått på hur stort felet är mellan facit och beräknad utdata, och denna förlust ligger till grund för uppdateringen av viktparametrarna (Chollet, 2017). Viktparametrar och biasparametrar måste uppdateras på ett effektivt sätt och för detta finns det olika funktioner (Chollet, 2017); hur dessa fungerar kommer inte att beskrivas ytterligare i det här arbetet.

För att en ANN-modell ska prestera så bra som möjligt bör indata först anpassas. Höga värden, och stora spann av värden, bör undvikas och därför bör indata korrigeras så att värden hamnar runt noll. Detta är giltigt i en ANN-modell eftersom det inte är de absoluta värdena som är viktiga.

Den ANN-modell som använts i det här arbetet är utvecklad av Rehn och Giertz (2019) på SVOA. Modellen är skriven i Python och utnyttjar modulen Keras, utvecklad av Chollet m. fl. (2015). Keras är skriven på ett sådant sätt att vid användning sker allt matematiskt i bakgrunden; användaren ansvarar endast för att förse modellen med indata samt beskriva hur modellen ska byggas upp och vilka metoder

som ska användas – exempelvis antalet lager och noder, hur uppdatering av vikter ska ske och vilka transferfunktioner som ska användas. ANN är ett samlingsnamn för ett flertal olika modeller. Den ANN-modell som utvecklats av David Rehn (Rehn & Giertz, 2019) och som användes i den här rapporten är en ”Multilayer perceptron” (MLP-modell), vilket är en typ av ANN-modell som består av minst tre lager.

2.2.1 Tillämpningar

ANN-modeller har använts i tidigare studier av både Kutyłowska (2016) och Jafar m. fl. (2010) för att utvärdera ledningsnät. Resultatet i båda studierna indikerar att ANN-modeller kan användas för att utvärdera ledningsnät.

Kutyłowska (2016) undersökte två olika sorters ANN-modeller, varav den ena var en MLP-ANN modell. Aktuella attribut i studien var antalet serviskopplingar, längder på huvudledningar, distributionsledningar och servisledningar samt fel på dessa ledningar.

Jafar m. fl. (2010) undersökte likt Kutyłowska (2016) renoveringsbehovet på vattenledningar med hjälp av en ANN-modell. Inparametrarna i Jafars studie delades in i tre huvudgrupper: fysiska faktorer, miljöfaktorer och operativa faktorer. I fysiska attribut inkluderades ledningens material, längd, diameter, tjocklek och ålder. I miljöfaktorer inräknades jordtyp och ledningsdjup och till operativa attribut räknades tryck och skydd. Utifrån korrelation mellan olika attribut delades ledningsobjekten in i undergrupper. Både Jafar m. fl. (2010) och Kutyłowska (2016) erhöll modeller med bra prestation.

2.2.2 Validering av en ANN-modell

Detta delavsnitt beskriver kortfattat hur ANN-modeller används vid modellering och faktorer som bör beaktas för att uppnå ett gott resultat.

Ett vanligt förfarande vid användande av en ANN-modell är att dela upp tillgänglig data i tre grupper: träningsdata, valideringsdata och testdata. Träningsdata är data som används för att träna modellen i att identifiera samband mellan indata och utdata. Valideringsdata är data

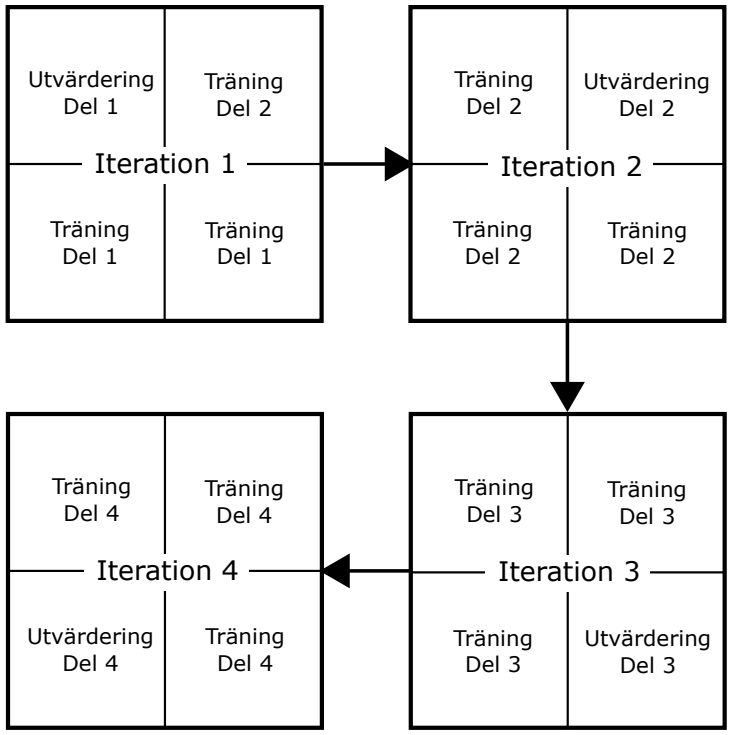
som används under träningsperioden för att utvärdera modellen. Detta är samma typ av data och kommer från samma datainsamling – skillnaden är att modellen inte tränas på valideringsdata, utan utvärderas bara. Ju fler epoker (iterationer) modellen tränas i, desto mindre kommer felet i utdata för träningsdata att vara. Efter ett visst antal iterationer kommer dock felet för valideringsdata att öka. Detta är, som beskrivits tidigare, ett resultat av överpassning (Chollet, 2017). Genom att identifiera det antal epoker där felet för valideringsdata är minst kan överpassning undvikas, eller effekten åtminstone minskas. En förutsättning för detta är att modellen aldrig tränas på valideringsdata. I praktiken kommer dock modellen att tränas med valideringsdata, eftersom valideringsdata används för att se hur modellen presterar och beroende på resultat kommer justeringar i modellen att göras. Detta benämns ”informationsläckage” (Chollet, 2017) och minskar kvaliteten på modellen. Därför är det viktigt att även ha testdata. Detta dataset är den avslutande kontrollen av modellen för att undersöka ifall den presterar väl och ska endast användas som en avslutande kontroll. Modellen får inte tränas på, eller valideras mot, testdata.

Att köra modellen på ett vattenledningsnät skiljer sig från att exempelvis klassa bilder på djur som hund eller inte hund. I det senare fallet används bilder som redan klassats till hund eller inte hund och ingen osäkerhet råder – i de allra flesta fall är det enkelt för en människa att säga att det är en hund eller inte. Modellen tränas därefter på dessa bilder för att lära sig karaktäristika som kännetecknar en hund. När modellen tränats färdigt körs modellen på testdata och ett slutgiltigt prestationsmått erhålls. Problemet med vattenledningar är att det finns oidentifierade läckor – det är därför modellen behövs. Historiska data är inte heller så omfattande eller väldokumenterad att ANN-modellen kan tränas på historiska data där det entydigt går att säga ifall en ledning hade en läcka eller inte. Modellen tränas därför på det befintliga ledningsnätet, det vill säga på samma ledningar som modellen ska användas på i ett senare skede. Detta innebär att facit innehåller fel – facit bygger på ledningarnas historik, och en ledning som haft minst en läcka klassas som ett, och ledningar där ingen läcka observerats klassas som noll, men vissa av ledningsobjekten klassade som noll har oidentifierade läckor.

Detta komplicerar användandet av testdata eftersom de prestationsmått som testkörningen genererar, exempelvis förlust och noggrannhet, inte kan tolkas lika entydigt som för hundbilder.

Det finns sätt att motverka både överpassning och informationsläckage i en ANN-modell, och sådana tekniker användes i den här studien. Överpassning kan motverkas genom att i modellen tilldela en kostnad för att använda stora vikter (*regularizers*), eller genom att slumpmässigt ta bort vissa attribut i indata (*dropout*) (Chollet, 2017). Kostnaden i *regularizers* utgörs av ett tillägg till modellens förlust baserat på viktternas storlek. I ANN-modellen som användes i den här studien prövades både dropout och regularizers och båda minskade överpassningen. Dropout används i Stockholmsmodellen (Rehn & Giertz, 2019), och därför användes det även i denna studie i slutänden.

Effekten av informationsläckage kan, enligt Chollet (2017), undvikas genom att använda itererad k-delad korsvalidering. En enklare variant av detta användes i det här arbetet: k-delad korsutvärdering, vilken genomförs genom att dela upp all data utöver testdata i k delar. Modellen tränas därefter på alla delar utom en del och valideras på den sista delen. Detta återupprepas k gånger och modellen tränas därför på alla delar. Figur 9 illustrerar hur k-delad korsutvärdering fungerar. Valideringens fel och noggrannhet är sedan medelvärdet av dessa, från de k körda modellerna (Chollet, 2017). Itererad k-delad korsvalidering är en utveckling på denna metod där datasetet blandas efter varje iteration så att unika delar uppstår varje gång. Den enklare metoden, k-delad korsutvärdering, är lättare att utföra och eftersom de båda metoderna bygger på samma förfarande bör även k-delad korsvalidering fungera för att minska risken för informationsläckage, om än inte lika väl. Anledningen till att den enklare metoden användes var för att Scikit-Learn (Pedregosa m. fl., 2011) möjliggör en enkel applicering av k-delad korsutvärdering. Egentligen är de båda metoderna utvecklade för att hantera problem som uppstår om antalet prover är få (Chollet, 2017), och detta är ytterligare en anledning till att k-delad korsutvärdering användes i det här arbetet.



Figur 9. En illustration av konceptet k-delad korsvalidering. Modellens fel och noggrannhet är medelvärdet av utfallen från de fyra iterationerna (omarbetat från Chollet (2017)).

2.3 Utvärdering av viktiga attribut

Detta avsnitt beskriver först varför det är en god idé att begränsa antalet attribut, för att sedan introducera existerande metoder för att utvärdera attribut på ett effektivt sätt. I det här arbetet användes främst en variant på Relief-algoritmen för attributurval: ReliefF. Hur grundalgoritmen Relief fungerar beskrivs i delavsnitt 2.3.1. Som komplement till ReliefF-algoritmen användes också två varianter av wrappermetoden Recursive Feature Elimination (RFE) och dessa beskrivs i delavsnitt 2.3.2. För att undersöka korrelationen mellan olika attribut utfördes också olika korrelationsanalyser, och de beskrivs i avsnittet delavsnitt 2.3.3. Alla metoder som användes i det här arbetet för attributurval sammanfattas i tabell 3.

Tabell 3. De attributurvalsmetoder som användes i det här arbetet. Vad respektive typ, familj och metod är förklaras i kommande delavsnitt.

Typ	Familj	Metod	Modul[1]
Filter	Relief	ReliefF	ReBATE[2]
	Korrelation	Cramérs V	-
	Korrelation	Spearmans rangkorrelation	pandas[3]
	Korrelation	Multinomial logistisk regression	Scikit-Learn[4]
Wrapper	RFE	Multinomial logistisk regression	Scikit-Learn[4]
	RFE	Random Forest Classification	Scikit-Learn[4]

[1] Alla är moduler till Python
[2] Urbanowicz, Olson m. fl. (2018)
[3] McKinney (2010) och The pandas development team (2020)
[4] Pedregosa m. fl. (2011)

Målet med denna studie var att utvärdera vilka attribut i det artificiella neuronnät som används i Ordning i RörANN som är särskilt viktiga för att identifiera vattenledningar med risk för läcka. Även om inga ändringar i modellen görs (exempelvis förändring i antalet lager och noder) finns det hundratusentals kombinationer av dessa attribut (baserat på antalet tillgängliga attribut för Umeå), och därför behövdes utvärderingen göras på ett strukturerat sätt.

En stor fördel med ANN-modeller är att de minskar behovet av att förbehandla data innan modellen körs. En viss bearbetning krävs fortfarande men inte till samma grad som andra maskininlärningsmetoder (Chollet, 2017; Salesi m. fl., 2018). Det kan dock finnas anledningar till att förbehandla data i större utsträckning (Robnik-Šikonja & Kononenko, 2003) och Salesi m. fl. (2018) fann i en studie att vid ett reducerat antal attribut i indata kunde ett bättre resultat uppnås. Även om slutresultatet skulle vara detsamma innebär färre attribut att beräkningstiden kommer att minska, vilket är fördelaktigt om antalet attribut är många (Kim m. fl., 2010; Urbanowicz, Meeker m. fl., 2018). Ytterligare ett problem med många attribut är att det ställer höga krav på mängden indata, vilket kan vara dyrt att erhålla (Sun, 2007), men även svårt att anskaffa.

Avseende underhållsbehov för vattenledningar innebär många attribut att det administrativa arbetet ökar om en stor mängd data måste samlas in och lagras. Ett annat problem är att ledningars livslängd är lång, och då ledningarna är förlagda i mark är det svårt att samla in information om ledningsnätet i efterhand. Om till exempel en lednings diameter är okänd, är det svårt att bestämma denna utan att gräva fram ledningen. Eftersom retroaktivt insamlande av data är komplicerat kan det vara svårt att applicera en ANN-modell på ett ledningsnät om inte ledningsinformation historiskt sett nedtecknats i hög grad. Det finns alltså flera anledningar till att begränsa antalet attribut i indata.

För att uppnå gott resultat även med en begränsad mängd attribut är det viktigt att de attribut som används är tillräckligt informationsrika för att modellen ska kunna uppskatta ledningars kondition på ett tillfredsställande sätt. Då attributurval varit ett aktuellt forskningsområde i flera år, finns det ett stort antal olika metoder för att identifiera viktiga attribut. Dessa kan grovt delas in i tre huvudgrupper: filtermetoder, wrappermetoder och inbäddade metoder (Hall, 1999; Thi & Nguyen, 2016). Filtermetoder använder olika tekniker för att identifiera viktiga attribut utan att involvera den faktiska modellen. Wrappermetoden optimerar i stället resultatet med hjälp av den faktiska modellen. De inbäddade metoderna är en kombination av filtermetodiken och wrappermetodiken, där attribut först väljs av en filtermetod för att sedan optimeras med hjälp av en wrappermetod (Thi & Nguyen, 2016).

Ett χ^2-test kan användas för attributurval och ingår i filtermetoder (Urbanowicz, Olson m. fl., 2018), och användes av Jafar m. fl. (2010) vid attributurval för att bestämma relevanta attribut för utvärdering av ett vattenledningsnät. ANOVA F-test är en annan statistisk metod som kan tillämpas för attributurval (Urbanowicz, Olson m. fl., 2018). En populär filtermetod är Relief (Kira & Randell, 1992) som med tiden vidareutvecklats till flera ny metoder (Robnik-Šikonja & Kononenko, 2003; Urbanowicz, Olson m. fl., 2018).

Wrappermetoder är iterativa metoder där en ANN-modell testkörs med olika attribut. Med stora prov där attributen är många, är det inte praktiskt möjligt att testa alla kombinationer. Därför tillämpar wrappermetoder olika tekniker för att testa kombinationer på ett effektivt

sätt. En metod, som benämns ”Incremental algorithms” (*algoritmer med stegvis ökning*), går ut på att börja med ett attribut för att därefter i varje ny iteration lägga till ett attribut tills kvaliteten slutar att förbättras (Castillo m. fl., 2000). En annan metod, ”Decremental algorithms” (*algoritmer med stegvis minskning*), gör motsatsen; den börjar med att inkludera alla attribut för att sedan ta bort ett attribut åt gången (Castillo m. fl., 2000). Att stegvis avtagande algoritmer börjar med att inkludera alla attribut får som konsekvens att de blir långsamma.

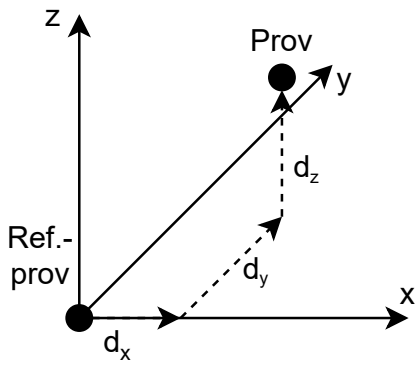
Filtermetoder är snabbare än wrappermetoder och kan uppnå, om än inte lika hög kvalitet som wrappermetoder, bra resultat (Hall, 1999). Filtermetoden ReliefF är den metod som främst användes i det här arbetet för attributurval. Valet gjordes eftersom ReliefF är snabbare än wrappermetoder, olika Relief-varianter har presterat väl i tidigare studier (Beretta & Santaniello, 2011; Salesi m. fl., 2018; Shi m. fl., 2017; Srinivas m. fl., 2019; Urbanowicz, Olson m. fl., 2018), och den är enkel att applicera i Python (Urbanowicz, Olson m. fl., 2018).

Som tidigare nämnts används träningsdata, valideringsdata och testdata när en ANN-modell tränas. Aldehim och Wang (2015) fann att all data kunde användas för attributurval – att inkludera testdata försämrade inte slutresultatet för den avslutande klassificeringsmodellen.

2.3.1 Relief-algoritmen

Relief baserar urvalet av viktiga attribut på närmsta granne-metoden. Förenklat innebär det att prov x jämförs med det prov som är mest likt prov x för att utvärdera viktiga attribut. Denna process kan sedan återupprepas tills alla prover undersökts. I varje iteration av Relief-algoritmen tilldelas de olika attributen en vikt beroende på ifall de beskriver utfallet på ett bra eller dåligt sätt (Robnik-Šikonja & Kononenko, 2003).

I detalj görs detta genom att ett prov först väljs slumpmässigt, i följande text benämnt referensprovet, varefter manhattanavståndet beräknas mellan referensprovet och resterande prover. Manhattanavståndet innebär att avståndet i respektive dimension mellan två prover beräknas och summeras, vilket illustreras i figur 10, där avstånden i respektive dimension är d_x, d_y och d_z och manhattanavståndet summan av dessa.



Figur 10. En illustration av närmsta granne-metoden. Origo representerar det slumpmässigt utvalda provet och d_x, d_y och d_z är avståndet mellan närmsta granne i tre dimensioner. Manhattanavståndet är summan av d_x, d_y och d_z.

När avståndet från referensprovet till resterande prover beräknats identifieras närmsta träff och miss. En träff är definierad som ett prov som har samma utfall som referensprovet, och en miss som ett prov som inte har samma utfall som referensprovet. För läcka på en vattenledning utgör en ledningssträcka ett prov, och en träff skulle innebära att både referensprovet och det andra provet indikerar, eller inte indikerar, läcka. När närmsta träff och miss identifierats, beräknas vikter för varje attribut. Ifall den *närmsta träffen* och referensprovet inte har samma värde för det aktuella attributet minskas vikten; om referensprovet och den *närmsta missen* inte delar värde ökas vikten. Hur vikten beräknas visas i ekvation 5.

$$W(A) = -\frac{\text{diff}(A, P_i, T)}{m} + \frac{\text{diff}(A, P_i, M)}{m} \tag{5}$$

W(A) är vikten för det aktuella attributet.

diff() är en funktion som beskrivs i ekvation 6 och 7.

(A,R_i,T) är attributet A som undersöks i den närmsta träffen T för provet P.

(A,R_i,M) är attributet A som undersöks i den närmsta missen M för provet P.

m är det antal gånger algoritmen ska köras.

Kononenko m. fl. (2000) föreslår att m sätts till det totala antalet prov och därefter väljs ett antal bästa attribut vilket medför att relationen mellan alla prov utnyttjas. En nackdel med detta är att beräkningstiden ökas.

För kategoriska attribut, det vill säga diskreta, är *diff* definierat som:

$$\text{diff}(A, P_1, P_2) = \begin{cases} 0; \text{ värde}(A, P_1) = \text{värde}(A, P_2) \\ 1; \text{ annars} \end{cases} \tag{6}$$

där P_1 är referensprovet och P_2 antingen är närmsta träff eller närmsta miss. För numeriska attribut, det vill säga kontinuerliga, beräknas diff i stället enligt:

$$\text{diff}(A, P_1, P_2) = \frac{|\text{värde}(A, P_1) - \text{värde}(A, P_2)|}{max(A) - min(A)} \tag{7}$$

Relief har en del begränsningar. Därför har olika försök till förbättringar av algoritmen gjorts. En tidig, och fortfarande populär, variant är ReliefF. ReliefF är mindre känslig för brus, kan hantera inkomplett data och kan hantera situationer där utfallet inte är binärt (Robnik-Šikonja & Kononenko, 2003). Till skillnad från Relief, tar ReliefF hänsyn till de k närmsta träffarna och missarna i stället för endast den närmsta träffen och missen.

För dataset där det kan finnas beroende mellan olika attribut ger ett lågt k bäst resultat: upp till en viss gräns ger fler k bättre uppskattning av attributs vikt, men när k ökas ytterligare inkluderas irrelevanta grannar och prestationen försämras – därför bör k inte vara alltför stort. När attribut är oberoende ökar algoritmens prestation med ökat k (Kononenko m. fl., 1997). ReliefF-vikterna går då mot den andel av totala utfall som respektive attribut är delaktig i att förklara. Hur många närmsta grannar som bör inkluderas förutsatt att beroende finns mellan attribut varierar med data (Robnik-Šikonja & Kononenko, 2003), men de tio närmsta grannarna rekommenderas av Kononenko m. fl. (1997).

Utöver ReliefF finns det ett flertal andra varianter på Relief. Urbanowicz, Olson m. fl. (2018) utförde ett utvärderingsprov (benchmark) på ett flertal attributurvalsmetoder där filtermetoder och wrappermetoder testades på olika dataset för att undersöka hur de presterade. Utvärde-

ringsprovet innefattade de ej Relief-baserade filtermetoderna χ^2, ANOVA F-test och Mutual information, de Relief-baserade filteralgoritmerna ReliefF, SURF, SURF*, MultiSURF* och MultiSURF samt wrappermetoderna ExtraTrees och REF ExtraTrees. Den slutsats som drogs var att de ej Relief-baserade algoritmerna och de två wrapperalgoritmerna presterade sämre än de olika Reliefbaserade algoritmerna.

En nackdel med Relief-baserade algoritmer är att om attributen både har kontinuerliga och diskreta värden kan betydelsen av de kontinuerliga värdena underskattas (Urbanowicz, Meeker m. fl., 2018). Detta kan motverkas med en rampfunktion, (Robnik-Šikonja & Kononenko, 2003) men denna metod introducerar två ytterligare variabler som måste bestämmas, vilket komplicerar metoden. Ett annat problem är att originalalgoritmen inte är bra på att identifiera om flera attribut tillsammans inducerar något; det vill säga att en grupp attribut kan leda till något som inget attribut i sig själv inducerar. Relief tar inte heller hänsyn till korrelation – två identiska attribut kommer därför få samma vikt trots att ett av attributen är överflödigt (Urbanowicz, Meeker m. fl., 2018). Vidare är algoritmen, som nämnts tidigare, känslig för brus, vilket dock de olika vidareutvecklingarna försökt hantera.

Till Pyhton har det utvecklats en modul som innehåller olika varianter av Relief-algoritmen (Urbanowicz, Olson m. fl., 2018). Denna modul distribueras under namnet ReBATE och modulen har använts i den här studien. Detta gör det enkelt att applicera Relief-algoritmer för attributurval men en nackdel är att mindre kontroll ges över algoritmen. Exempelvis kan det vara svårt att applicera tidigare nämnd rampfunktion för att minska underskattningen av kontinuerliga attribut.

2.3.2 Wrappers

Wrappermetoder väljer ut viktiga attribut i en modell genom att köra modellen upprepade gånger, och, beroende på typ av wrapper, lägger till eller tar bort attribut i varje iteration baserat på modellens prestation. Eftersom ANN-modeller inte ger något mått på hur viktiga olika attribut är efter en körning är det svårt att använda wrappers för ANN-modeller. Det finns dock andra modeller som ger mått på hur viktiga olika attribut

är efter varje körning. Dessa mått på ett attributs vikt kan då utnyttjas för att undersöka olika attributs relevans med en wrappermetod.

En förekommande wrappermetodik är Recursive Feature Elimination, RFE (ungefär återupprepande attributseliminering). Detta är en wrapper där olika attributs relevans först rangordnas. Därefter körs modellen och utvärderas, och de lägst rangordnade attributen elimineras. Därefter rangordnas attributen igen och modellen tränas och utvärderas återigen. Detta återupprepas tills inga attribut finns kvar (Gregorutti m. fl., 2016). RFE har i flera studier uppvisat gott resultat (Brungard m. fl., 2015; Gregorutti m. fl., 2016; Guyon m. fl., 2002). De olika studierna har utnyttjat olika klassificeringsalgoritmer: Guyon m. fl. (2002) utvecklade RFE och använde metoden med algoritmen Stödvektormaskin, Gregorutti m. fl. (2016) använde den tillsammans med en Random Forest-algoritm och Brungard m. fl. (2015) använde algoritmen tillsammans med flera olika klassificeringsalgoritmer i syfte att jämföra algoritmerna.

RFE tillsammans med Random Forest-klassificering (RFC) har uppvisat gott resultat i tidigare studier (Brungard m. fl., 2015; Gregorutti m. fl., 2016). RFC kan sägas vara en utveckling av metoden Decision Tree (Nisbet m. fl., 2018), där klassificering i Decision Tree genomförs genom att stegvis separera indata i undergrupper, med hjälp av olika kriterier. RFC skapar i stället ett stort antal träd där undergrupper skapas genom slumpmässiga regler. Den slutgiltiga klassificeringen är sedan medelvärdet av alla träds utfall (Gregorutti m. fl., 2016). Brungard m. fl. (2015) klassade olika algoritmer att köra tillsammans med RFE i klasserna enkla, medelsvåra och komplicerade att använda. RFC klassades som komplicerad, och ett alternativ till RFC, med klassningen enkel, var multinomial logistisk regressionsanalys (MLR). MLR beskrivs mer noggrant i avsnitt 2.3.3 och kommer tillsammans med RFE användas som komplement till RFC i den här studien – att få ett bra resultat från RFC kan vara mer komplicerat än att få ett bra resultat av multinomial logistisk regression, även om en väl genomförd RFC kan ge ett bättre resultat. De båda wrapperanalyserna gjordes med hjälp av Scikit-Learn (Pedregosa m. fl., 2011).

2.3.3 Korrelation

För att undersöka korrelation mellan attribut kan olika korrelationsanalyser genomföras. Vilken korrelationsanalys som är lämplig skiljer sig beroende på ifall de attribut som ska utvärderas är av nominal, ordinal, intervall- eller kvottyp. Därför användes tre olika korrelationstester i den här studien.

För kategoriska attribut (inkluderar nominala och ordinala värden) kan Cramérs V användas. Cramérs V normaliserar χ^2-testet så att korrelationsvärdet varierar mellan noll och ett, där ett innebär fullständig korrelation (Acock & Stavig, 1979; Bergsma, 2012) och beräknas enligt ekvation 8 (Bergsma, 2012).

$$V = \sqrt{\frac{\phi}{min(r-1,\ c-1)}} \tag{8}$$

Cramérs V är dock betungad av bias och Bergsma (2012) föreslår därför en korrigering för detta, redovisad i ekvation 9.

$$\begin{aligned} \tilde{\phi}^2 &= \hat{\phi}^2 - \frac{1}{n-1}(r-1)(c-1) \\ \hat{\phi}_+^2 &= max(0,\ \tilde{\phi}^2) \\ \tilde{r} &= r - \frac{1}{n-1}(r-1)^2 \\ \tilde{c} &= c - \frac{1}{n-1}(c-1)^2 \\ \tilde{V} &= \sqrt{\frac{\phi_+^2}{min(\tilde{r}-1, \tilde{c}-1)}} \end{aligned} \tag{9}$$

där

$\hat{\phi}$ är observerad χ^2

$\tilde{\phi}$ är χ^2-värdet korrigerad för bias

$\tilde{r}$ är antalet unika attribut för det ena attributet

$\tilde{c}$ är antalet unika attribut för det andra attributet

$\tilde{V}$ är det korrigerade observerade Cramérs V.

En förutsättning för att χ^2-testet ska ge ett bra resultat är att endast 20 procent av cellerna i kontingenstabellen mellan två attribut har ett värde under 5 (SPSS, 2020). En kontingenstabell är en tabell där två attribut jämförs och där det för varje kombination av kategorier räknas hur många prov som har den kombinationen. Detta exemplifieras i tabell 4. Två egenskaper jämförs: klädesplagg och färg. Tabellen visar att fem prover hade blå tröja, tio prover hade blå byxor och så vidare. Förutsättningen för χ^2-testet är alltså att endast ett begränsat antal av dessa kombinationer får ha ett värde under fem.

Tabell 4. Ett exempel på en kontingenstabell.

	Färg			
Klädesplagg	Blå	Grön	Grå	Svart
Tröja	5	2	3	10
Byxa	10	0	5	5

Om residualerna från data är normalfördelade eller ej påverkar vilka korrelationsanalyser som är lämpliga för kontinuerliga attribut. För att undersöka ifall residualerna är normalfördelade kan ett Jarque-Bera test användas (MathWorks, 2020; Seabold & Perktold, 2010). Om ett stort JB-värde erhålls förkastas nollhypotesen att data är normalfördelade.

För numeriska attribut kan Pearsons korrelationskoefficient användas och beräknas enligt ekvation 10 (Rodgers & Nicewander, 1988). Pearson korrelationskoefficient förutsätter att residualerna är normalfördelade (Studenmund, 2011) och om så inte är fallet kan Spearmans rangkorrelation användas (McDonald, 2014). De kontinuerliga attributens residualer visade sig inte vara normalfördelade (se appendix 2.3.3) och därför användes Spearmans rangkorrelation. Dessutom undersöker Pearsons korrelationstest endast linjär korrelation – Spearmans rangkorrelation undersöker i stället monoton korrelation, det vill säga ifall den beroende variablen ökar eller minskar när den oberoende variabeln ökar. ANN-modellen kan identifiera olinjära samband och detta är ytterligare en anledning till att Spearmans rangkorrelation användes. Spearmans

rangkorrelation är en anpassning av Pearsons korrelationskoefficient och beräknas genom att först rangordna alla värden, och därefter beräkna Pearsons korrelationskoefficient på de rangordnade värdena.

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\left(\sum(X_i - \bar{X})^2 + \sum(Y_i - \bar{Y})^2\right)^{1/2}} \tag{10}$$

där

$\bar{X}$ är medelvärdet för attribut X

$\bar{Y}$ är medelvärdet för attribut Y

X_i är värdet för den i:te förekomsten av X

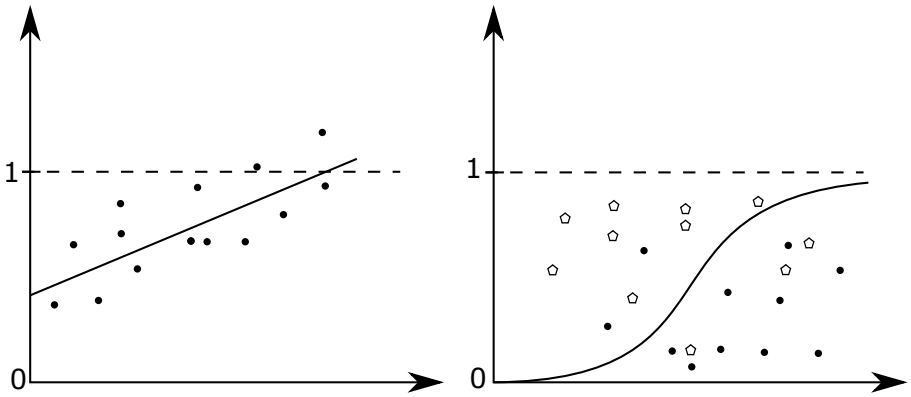
Y_i är värdet för den i:te förekomsten av Y

Kombinationen kategoriska och numeriska attribut är svårare, speciellt ifall de kategoriska attributen är nominala. Ett sätt att hantera detta är genom multinomial logistisk regression. För att förstå MLR är det enklare att börja med linjär regression. Vid en enkel linjär regression baserad på minstakvadratmetoden anpassas en linje så att kvadratfelet blir så litet som möjligt, se ekvation 11, där y_i är uppmätt värde, och $f(x_i)$ är beräknat värde för prov i.

$$\frac{d}{dx}\left(\sum\left(y_i - f(x_i)\right)^2\right) = 0 \tag{11}$$

Om den beroende variabeln endast kan ta värdena noll eller ett blir tolkningen av resultatet från en linjär regression svår. $f(x) > 0,5$ kan klassas som 1, och $f(x) < 0,5$ till 0, men med en linjär funktion kan $f(x) > 1$, vilket inte har en riktig innebörd. Därför kan enkel logistisk regression användas. Då omvandlas värdena så att en logistisk funktion bildas, vilken går mot 0 då $x \to -\infty$ och 1 då $x \to \infty$ (Studenmund, 2011). I stället för att försöka passa en kurva så väl som möjligt till punkter är logistisk regression en iterativ process där olika kurvor testas, och i varje iteration undersöks kurvans förmåga att dela in punkterna i rätt klasser. Data i det här arbetet innehåller attribut med fler än två kategorier och då fungerar inte enkel logistisk regression. I stället kan

multonomial logistisk regression, MLR, användas (Studenmund, 2011). Den exakta teorin bakom denna metod kommer inte beröras, men i MLR kan den beroende variabeln ha fler utfall än två. Skillnaden mellan linjär regression och logistisk regression illustreras i figur 11.



Figur 11. Den vänstra bilden är ett exempel på en minstakvadratsregression. Den högra bilden är ett exempel på logistisk regression.
Minstakvadratsregression anpassar en linje till punkter. Logistisk regression anpassar en kurva för att separera två klasser.

Ett mått på en regressions förmåga att separera klasser korrekt är R-värdet, definierat som antalet korrekt klassade värdena dividerat med det totala antalet värden. Detta innebär dock att prestationsmåttet blir missvisande om ojämnhet i indata förekommer. Om exempelvis 80 procent av indata har klass 0 och 20 procent klass 1, kommer R-värdet bli 0,8 om alla värden klassas till 0, vilket vid en första anblick antyder att modellen gör ett bra jobb förutspå klasserna, fastän modellen endast klassar alla värden till 0. Ett annat mått, $\bar{R}_p$, föreslås av Studenmund (2011) och definieras enligt ekvation 12 för två kategorier. Detta mått användes i den här studien då fördelningen mellan olika klasser inte var homogen.

$$\bar{R}_p = \frac{\frac{n_{i_{korr}}}{n_i} + \frac{n_{j_{korr}}}{n_j}}{2} \tag{12}$$

där

n_i är antalet prov av klass i

n_j är antalet prov av klass j

$n_{i_{klass}}$ är antalet korrekt klassificerade prov av klass i

$n_{j_{klass}}$ är antalet korrekt klassificerade prov av klass j

3 Metod

Delavsnitt 3.1 beskriver överväganden och förenklingar som gjordes i datainsamlandet. En mer utförlig beskrivning av hur data samlades in för respektive attribut finns i appendix A. Delavsnitt 3.2 redovisar hur data bearbetades för att kunna användas i ANN-modellen. De två efterföljande delavsnitten, 3.3 och 3.4, beskriver hur modellen tränades, respektive hur attributen utvärderades. Avsnitt 3.5 behandlar hur prediktioner från den körda ANN-modellen användes för att analysera bidragande faktorer till dålig kondition på ledningar.

3.1 Insamlande av data

För att genomföra utvärderingen av attribut i ANN-modellen behövdes först data för Umeå kommun sammanställas, och som inspiration användes Stockholmsmodellen (Rehn & Giertz, 2019). För vissa av de attribut som användes i Stockholm saknades det data för i Umeå, och dessa attribut sållades bort direkt. Dessutom har vissa attribut lagts till som inte förekom i Stockholmsmodellen. De attribut som användes i Umeåmodellen respektive Stockholmsmodellen redovisas i tabell 5. Nämnas bör att det kan förekomma skillnader i hur de olika attributen inkorporerats i respektive modell. Nästan inga fysikaliska attribut har använts. Det hade exempelvis varit intressant att ha med internt och externt tryck, men data har inte funnits för att kunna ha med sådana attribut. Fokus har legat på attribut som Umeå redan hade tillgång till på ett eller annat sätt för att göra modellen så enkel som möjligt att använda.

Tabell 5. Vilka attribut som används i Umeåmodellen respektive Stockholmsmodellen.

Attribut	Umeå	Stockholm
Anläggningsår	X	X
Avstånd från huvudledning	-	X
Befolkningsförändring	X	X
Bergtyp	X	-
Brandpostkoppling	-	X
Dimensionsändring	X	X
Driftstörningar	-	X
Fjärrvärme	X	-
Geologi	-	X
Höjd	X	X
Innerdiameter	X	X
Jordtyp	X	-
Järnväg	X	X
Korrosionsskydd	-	X
Markanvändning	X	X
Material	X	X
Meteorologi	-	X
Relinad	X	-
Renovering	-	X
Serviskoppling	X	X
Stadsdel	X	-
Trafiklast	X	X
Tryckzon	X	X
Ventilkoppling	X	X
Ålder	X	X

All data som användes var lagrade i olika kartlager, och relevant information extraherades ur dessa med hjälp av två olika GIS-programvaror: QGIS och VA-banken. Vattenledningssystemet var lagrat som linjer, och attribut knöts till dessa med hjälp av överlagringsfunktioner där data från ett kartlager kopplas till ett annat kartlager baserat på spatiala relationer. I andra fall behövdes inga rumsliga relationer utnyttjas, utan information kunde extraheras direkt från kartlager, såsom ledningsdimensioner. I

tabell A1 i appendix A sammanfattas vilka operationer som använts för att extrahera data för de olika attributen.

Innerdiameter fanns inte angivet för alla ledningar, och för PE-ledningar var alltid endast ytterdiametern angiven. De ledningar som inte var av PE och saknade innerdiameter uteslöts då de dels var få, dels saknades en bra metod för att bestämma innerdiametern för de ledningarna. För PE-ledningar användes produktdata från Pipelife[5] för att bestämma innerdiameter. Detta förfarande medför en viss osäkerhet, då godstjockleken kan ha varierat med tiden, men eftersom innerdiameter kan antas vara relevant, och PE är ett populärt ledningsmaterial, ansågs denna metod ändå vara bättre än alternativet att ignorera alla PE-ledningar.

För att inkorporera järnväg behövdes det bestämmas vilka ledningar som kunde anses vara påverkade av järnvägen och dess trafik. Eftersom laster, vibrationer och så vidare sprider sig, bestämdes ett avstånd vari ledningar kunde anses vara påverkade. En ledning ansågs vara påverkad av järnväg om den låg inom tio meter från en järnväg, och dessa ledningar tilldelades en etta, de andra en nolla.

Både anläggningsår och ålder togs med eftersom de nödvändigtvis inte innebär samma sak, även om ålder är direkt beroende av anläggningsår. Som ett exempel kan tidigare nämnda Ehri-muffen användas (se avsnitt 2.1.1). Att rör med Ehri-muffar fallerar beror på en bristande konstruktion och är beroende av anläggningsår, inte ålder.

Geologi användes i Stockholmsmodellen (Rehn & Giertz, 2019), men var svårt att inkorporera eftersom geologi är ett brett begrepp, och kan innefatta jordart, bergart, förkastningar, strukturer och så vidare. I den här studien beaktades de två faktorerna jord- och bergart. Jordart var svårt att ta hänsyn till på ett bra sätt då flera olika kartlager fanns tillgängliga: djuplager, grundlager, ytlager och översta ytlager. Det var inte givet vilket lager som skulle användas – vilket lager ligger ledningen i, är det omkringliggande lager som är viktigast eller är det underliggande och så vidare? Dessutom kan fyllnadsmaterialet vara något helt annat. Grundlagret beskrivs av SGU som: ”Lagret avser den jordartstyp som normalt kan förväntas på karteringsdjup, dvs. ca 0,5 m under markytan,

[5]https://www.pipelife.se/se/

och som bedömas ha en mäktighet väl överstigande 0,5 meter."[6] Men den typ av grundlager som fanns tillgänglig för Umeå var av typen JG2, vilket SGU beskriver enligt: "Kartbilden är mycket kraftigt generaliserad och jordartsindelningen är grovt förenklad. Minsta redovisade yta är ca 1 km^2."[7] De avråder också från att använda kartlagret JG2 för analyser. Baserat på utredningen bör grundlagret ändå vara mest lämpligt och JG2 användes likväl för analys i det här arbetet. I grundlagret ingick totalt 17 olika klasser; för att konsolidera informationen, och för att ta hänsyn till osäkerheten, delades de olika jordarterna in i åtta grupper. Hur indelningen gjordes redovisas i tabell 6.

Tabell 6. Tabellen visar i vilka grupper de olika jordarterna delades in i.

Jordart	Grupp
Berg	Berg
Talus	Grovt
Klapper	
Morän	Morän
Isälvssediment, sand	Friktionsjord
Postglacial grovsilt – finsand	
Älvsediment grovsilt – finsand	
Isälvssediment	
Älvsediment, sand	
Svallsediment, grus	
Postglacial sand	
Flygsand	
Fyllning	
Lera – Silt	Kohesionsjord
Torv	Torv
Vatten	Vatten
Oklassat område	Oklassat

Kartlagret beskrivande markanvändning bestod av totalt 20 olika

[6] https://resource.sgu.se/dokument/produkter/jordarter-25-100000-wms-beskrivning.pdf

[7] https://resource.sgu.se/dokument/produkter/jordarter-1miljon-beskrivning.pdf

klasser. För att minska komplexiteten i modellen samlandes liknande markanvändning i grupper. Antalet klasser reducerades då till sju stycken. Vilka grupper markanvändning grupperades i redovisas i tabell 7.

Tabell 7. Hur de olika markanvändningsklassificeringarna delades in i nya klasser.

Ursprungsklass	Ny klass
Land utan nuvarande användning	Grönytor
Gröna urbana områden	
Jordbruk, semi-naturliga områden och våtmarker	
Kont. urban miljö >80 %	Urban miljö
Ej kont. urban miljö 50–80 %	
Ej kont. urban miljö 30–50 %	
Ej kont. urban miljö 10–50 %	
Ej kont. urban miljö <10 %	
Isolerade strukturer	
Sport- och fritidsfaciliteter	
Flygplatser	Industrimark
Byggarbetsplatser	
Industriella och kommersiella områden	
Mineralutvinning och avfallsstationer	
Hamnar	
Skog	Skog
Vägar och tillhörande mark	Vägar och tillhörande mark
Järnvägar och tillhörande mark	Järnvägar och tillhörande mark
Vatten	Vatten

Befolkningsutveckling hanterades genom att utifrån befolkningsstatistik klassa Umeås stadsdelar som ökande, konstant respektive minskande befolkningsmängd. Denna information överlagrades sedan på ledningsnätet. En noggrannare beskrivning ges i i tabell A1. Vidare gjordes vissa anpassningar för att stadsdelar bättre skulle passa vattenledningarnas utbredning för att undvika korta sträckor av vattenledningar

i angränsande stadsdelar.

För att koppla höjd till ledningsnätet användes rasterdata som överlagrades på dricksvattennätet genom en överlagringsfunktion i QGIS som kunde hantera en kombination av raster- och vektordata. Både den minimala och maximala höjden för ett ledningsobjekt behölls för att eventuellt fånga ett samband mellan lutning på en ledning och kondition.

3.1.1 Vindeln

Ett delmål av det här arbetet var att utvärdera ifall det var möjligt att köra ANN-modellen på en ort med sämre datatillgång och till denna fallstudie användes Vindelns kommun. I samband med detta används orden *enkla* och *generella*. Med enkla attribut åsyftas attribut som många kommuner bör ha tillgång till, eller som de kan få fram eller uppskatta. Generella attribut syftar på attribut som är desamma mellan kommuner: stadsdelar skiljer sig mellan kommuner, men rörmaterial bör oftast överensstämma. ANN-modellen kördes inte på Vindeln, men Umeå-modellen tränades och utvärderades med attribut som skulle kunna vara tillgängliga för Vindeln, och dessa attribut redovisas i tabell 8. Urvalet baserades dels på lärdomar från den datainsamling som gjordes för Umeå kommun, dels via kommunikation med sakkunnig på Vakin[8]. Anläggningsår och material är inte dokumenterat utförligt i Vindeln, men anläggningsår skulle kunna uppskattas genom att ta reda på när områden byggdes ut, och material kan eventuellt uppskattas utifrån anläggningsår.

[8]Petter Walan, Utredningsingenjör på Vakin

Tabell 8. Attribut som bör gå att få fram för Vindeln. Dessa ligger som grund för enkla och generella attribut.

Attribut tillgängliga i Vindeln
Anläggningsår
Bergtyp
Dimensionsförändring
Innerdiameter
Jordtyp
Markanvändning
Material
Maxhöjd
Minimihöjd
Närliggande järnväg
Serviskoppling
Trafiklast
Ventilkoppling
Ålder

3.2 Manipulering av data

Insamlad data behövde manipuleras för att kunna användas i ANN-modellen. Först och främst rensades vissa data bort. ANN-modeller kan hantera avsaknad av data så länge värdena kan ges ett värde som inte representerar något. För de kategoriska numeriska attributen, såsom antal serviser och närliggande järnväg, var det svårt att ge avsaknad av värde ett icke-betydande värde, och därför togs prover där värden saknades för dessa attribut bort – hade fler värden saknats hade det varit aktuellt med andra metoder för att hantera avsaknad av värden.

För de kontinuerliga attributen var det svårt att tilldela ett värde som entydigt kunde beskriva avsaknad av värde; ett värde som kan användas är noll men det är inte alltid lämpligt eftersom attribut kan ha värden som är, eller är nära, noll. Värden bör inte heller ta stora värden eller vara inom ett stort spann, varpå data ofta skalas om. Det är inte effektivt att skala om data ifall uteliggare förekommer; ponera att ett dataset har värdena 1, 1000, 1002 och 1003, skalas dessa värden om kommer 1 att

hamna nära 0 och 1000, 1002 och 1003 nära 1 vilket inte är önskvärt eftersom majoriteten av intervallet [0, 1] inte säger någonting. Med anledningen av detta är det inte lämpligt att tilldela avsaknad av värden ett godtyckligt värde såsom -999, även om det till skillnad från 0 inte är lätt att förväxla. Ledningsobjekt utan anläggningsår uteslöts eftersom värdet 0 skulle leda till ett stort spann. Det är möjligt att ett värde såsom 1850 skulle kunna ha tilldelats. Ledningsobjekt utan höjd uteslöts eftersom höjd kan vara nära, eller vara noll meter över havet. Däremot tilldelades ledningsobjekt utan innerdiameter nollor för avsaknad av värde eftersom spannet för innerdiameter inte påverkas nämnvärt ifall värdet noll introduceras, och innerdiametern för ett ledningsobjekt är aldrig naturligt noll. För att skala om de kontinuerliga attributens spann användes Scikit-Learn (Pedregosa m. fl., 2011).

De ej numeriska kategoriska dataseten omvandlades till dataset med dummy-variabler för att kunna hanteras i modellen. Detta kan exemplifieras med hjälp av figur 12. Totalt finns det tre kategorier: kategori 1, 2 och 3. Attribut A omvandlas till tre nya attribut – B, C och D – vilket motsvarar det totala antalet unika kategorier, där attribut B motsvarar kategori 1 och så vidare. Prov 1 tillhör kategori 1, och med dummy-variabler representeras det med en etta för attribut B och nollor för resterande attribut. Med samma resonemang tillhör prov 2 och 4 kategori 2 och så vidare. I prov 5 saknas värde, och detta representeras med att alla attribut blir 0.

Ett problem med den här metoden är att den kan öka dimensionaliteten i indata markant. När alla attribut i den här studien togs med utan dummy-variabler fanns 19 attribut. När kategoriska attribut omvandlats till dummy-variabler i enlighet med ovan återfanns 58 attribut, det vill säga mer än tre gånger så många attribut.

Efter dessa operationer kvarstod 9 440 unika ledningsobjekt utav 11713, varav 792 var klassade som ettor – det vill säga observerad läcka.

3.3 Modellen

Som beskrivits i avsnitt 2.2.2 innehåller facit fel vilket försvårar tolkningen av testdata. Mängden indata var också låg och därför prioriterades mer

	Attribut
Prov	A
1	Kategori 1
2	Kategori 2
3	Kategori 3
4	Kategori 2
5	Saknas

Omvandling till dummy-variabler

	Attribut		
Prov	B	C	D
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0
5	0	0	0

Figur 12. Figuren visar hur ett attribut med kategorier representeras med dummy-variabler.

indata i stället för användandet av testdata. Detta är inte oproblematiskt och kommer behandlas ytterligare i diskussionen.

Eftersom modellen tränades på de ledningsobjekt den ska utvärdera, tränades modellen på en begränsad mängd indata. Denna indata utgjordes av alla ledningar klassade som ett, samt ett slumpmässigt urval av lika många ledningar klassade som noll. Förhoppningen med modellen är att den ska lära sig att identifiera ledningar med läckor på den begränsade mängden data, för att sedan lyckas identifiera läckor på de ledningsobjekt som modellen inte sett. Tränas modellen på alla objekt kommer modellen lära sig att det inte är fel på de ledningar som är klassade som noll, fastän det i verkligheten kan vara så att de ledningarna har oidentifierade läckor. Eftersom modellen tränades på lika delar ettor som nollor tränades modellen på 1 584 ledningsobjekt varav 792 var klassade som ettor. Med denna konfiguration var det svårt att få modellen att prestera väl. Därför duplicerades ledningarna klassade som ettor, varpå 1 584 ledningar med läckor erhölls och modellen tränades då på totalt 3 168 ledningar. Noggrannheten blev då mycket bättre. Att duplicera värdena innebar vissa problem. Ju fler ledningsobjekt modellen tränades på, desto färre ledningsobjekt fanns kvar att utvärdera som modellen inte redan beaktat i träningsfasen. Att duplicera ledningsobjekten ytterligare hade eventuellt kunnat leda till en modell med högre noggrannhet, men eftersom idén är att träna modellen på en begränsad mängd data för att sedan utvärdera data för hela ledningsnätet måste det göras en avvägning mellan modellens prestation och vad som är syftet med modellen. Eftersom facit innehåller

fel är det problematiskt att låta modellen se alla värden. Därför ansågs en dubblering av antalet läckor vara en bra kompromiss – noggrannheten ökades markant, och majoriteten av ledningsobjekten användes inte för att träna modellen.

Arbetets mål var att utvärdera de viktigaste attributen vid körning av en ANN-modell på vattenledningsnät. För att kunna utvärdera modellens prestation med olika attribut behövdes först ett mått på hur bra modellen kunde prestera. ANN-modellen utvecklades därför för att prestera väl då alla attribut användes. Ursprungsmodellens baserades på Stockholmsmodellen (Rehn & Giertz, 2019).

3.4 Utvärdering av attribut

Totalt användes tre övergripande metoder för att utvärdera viktiga attribut: ReliefF, Recursive Feature Elimination (RFE) och korrelationsanalys.

I ReliefF-algoritmen uttrycktes nominala attribut i dummy-variabler vilket medförde att en vikt för varje dummy-variabel erhölls. De olika dummy-variablerna för ett attribut slogs ihop, och i varje steg behölls den högsta vikten. Exempelvis om *Attribut* är en nominal variabel med två kategorier omvandlas attributet till två dummy variabler: attribut_1 och attribut_2. Om attribut_1 och attribut_2 har vikterna 0,2 respektive 0,3 vid k=1, och 0,5 respektive 0,2 vid k=2 kommer den sammanslagna variabeln *Attribut* ha vikterna 0,3 och 0,5 vid k=1 respektive k=2.

Eftersom ReliefF inte kan identifiera attribut som ger liknande information undersöktes korrelation mellan de olika attributen. Attributen var av både kontinuerlig och kategorisk typ och därför behövdes totalt tre olika metoder för att analysera korrelation användas. Spearmans rangkorrelation för kontinuerliga värden, Cramérs V för kategoriska värden och MLR för kombinationen av kontinuerliga och kategoriska värden. Spearmans rangkorrelation användes i stället för Pearsons korrelationskoefficient eftersom kontinuerliga data inte var normalfördelade (se appendix 2.3.3). För att kunna jämföra korrelationen med ett gemensamt mått gjordes de kontinuerliga attributen, bortsett från höjderna, om till kategoriska attribut i analysen med Cramérs V. Anläggningsår och Ålder delades upp i fem intervall, där intervallen inom respektive

attribut var lika stora. Innerdiameter delades in i fem intervall där respektive intervall utformades för att fånga ett visst användningsområde för ledningarna. Minimi- och Maxhöjd gjordes inte om till kategoriska attribut då dessa två attribut bröt mot riktlinjen att max 20 procent av cellerna i kontingenstabell vid ett χ^2-test får understiga värdet 5.

Vid korrelationstestet med Cramérs V var det inte möjligt att undersöka alla kategoriska attribut tillsammans. I den kontingenstabell som bildas vid beräknandet av χ^2-värdet, som är en en förutsättning för Cramérs V, får endast 20 procent av cellerna ha ett värde under 5. Både Stadsdel och Närliggande järnväg bröt mot detta kriterium vid jämförelse med vissa attribut. I de fallen gjordes ingen korrelationsanalys mellan attributen. En lösning för Stadsdel hade varit att dela in Umeå i färre områden, men för att kunna fånga detaljer mellan stadsdelar måste ändå en tydlig uppdelning göras, och även med 20 områden blev det en grov indelning.

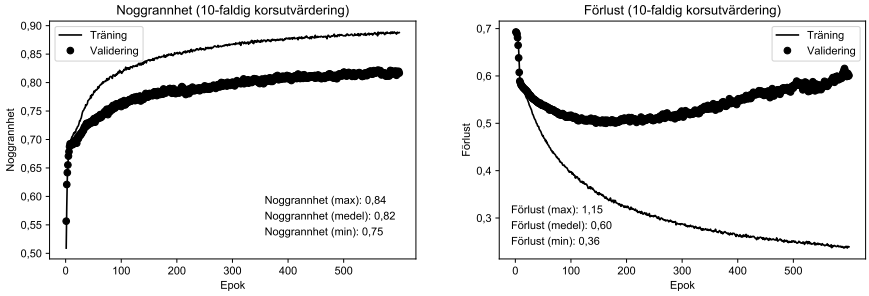
Som alternativ till ReliefF användes wrappermetodiken RFE, både med RFC och MLR. I RFE-algoritmen inkluderades också 10-faldig korsutvärdering för att kunna träna modellen på en större mängd data. De olika wreppermetoderna applicerades med hjälp av Pedregosa m. fl. (2011). För att köra analyserna gjordes nominala attribut om till dummy-variabler. Eftersom beräkningstiden var kort uteslöts endast ett attribut åt gången. Att de nominala attributen omvandlades till dummy-variabler enligt figur 12 medförde att varje dummy-variabel sågs som ett eget attribut och detta kan ha försvårat för wrappermetoderna att identifiera viktiga attribut. Detsamma gäller för ReliefF-algoritmen. För RFC kan dummy-variabler hanteras på ett annat sätt, men någon sådan funktion är ännu inte tillgänglig i Scikit-Learn. Scikit-Learn användes eftersom det är en populär katalog av funktioner som är lätta att använda.

Resultaten från de olika analyserna användes sedan för att forma delmängder av de tillgängliga attributen för att utvärdera modellen med dessa. Utöver detta utvärderades också modellen genom att kombinera attribut som hade stor påverkan på modellens prestation. Kombinationer av attribut som bör vara enkla att erhålla och generalisera prövades också.

För att begränsa antalet kombinationer varierades endast antalet noder

och epoker för modellen – ingenting annat. Antalet noder förändrades eftersom ju fler attribut som inkluderas, desto mer måste modellen lära sig, och ju fler noder modellen har, desto mer komplicerade samband kan modellen lära sig. Risken med för många noder är att modellen kan överpassas och därför kördes delmängder med få attribut både med många och få noder för att se skillnaden. Resultatet med få noder ansågs dock vara mer trovärdigt än det med många noder i och med risken för överpassning.

Antalet epoker varierades eftersom antalet noder och attribut påverkar när modellen börjar överpassa. Detta kan exemplifieras med figur 13. Detta är inte en körning som använts i resultatet, utan används för att visa hur antalet epoker kan justeras för att minska överpassning. Vid ungefär 150 epoker slutar förlusten att minska och ökar i stället. Detta innebär att felet efter varje ny epok blir större och större. Detta är ett tecken på överpassning, och för att minska överpassningen kan modellen köras igen fast med endast 150 epoker. Tillsynes motsägelsefullt är att noggrannheten fortsätter att öka, men det kan bero på att prediktionerna från den sigmoida funktionen $f(x)$ kommer längre och längre bort från de optimala värdena på 0 och 1, samtidigt som fler ledningar ändå hamnar på rätt sida om 0,5-sträcket och därför ändå klassas rätt, vilket påverkar noggrannheten positivt..



Figur 13. Figuren visar hur en grafer över modellens noggrannhet och förlust kan användas för att anpassa antalet epoker.

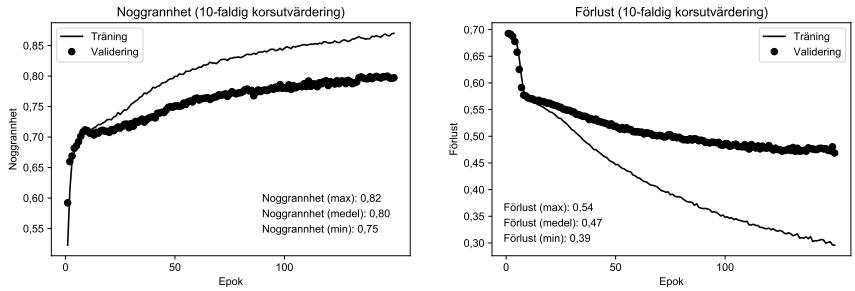
3.5 Analys av resultat

Utifrån de prediktioner av ledningsnätets kvalitet som erhölls efter att modellen körts kunde faktorers påverkan på ledningsnätet analyseras. För att utvärdera faktorers påverkan på ledningskvaliteten plottades attribut mot antalet läckor per hundra meter. Eftersom faktorers påverkansgrad varierar med ledningsmaterial, grupperades vissa attribut på ledningsmaterial innan data plottades. Därefter jämfördes resultatet med tidigare studier för att utvärdera ifall utfallet av ANN-modellen verkade stämma överens med vad som observerats tidigare eller inte.

4 Resultat

I det här avsnittet beskrivs först resultatet från de olika typer av attributurvalsmetoder som användes. Attributurvalsmetoderna är sammanfattade i tabell 3. Hur dessa sedermera användes för attributurval beskrivs i delavsnitt 4.4. Därefter redovisas en analys av olika attributs påverkan på läckfrekvens och totala antalet läckor på ledningsnätet i delavsnitt 4.6.

När modellen kördes med alla attribut uppnåddes en noggrannhet på 0,80 och en förlust på 0,47 (figur 14). Modellen kördes i 150 epoker. Detta är den körning av modellen som ligger till grund för utvärderingen av Umeås ledningsnät. Det är *inte* samma körning som modell 19 (tabell C1) som senare i avsnittet diskuteras i samband med attributurval – båda dessa körningar inkluderar dock alla attribut.

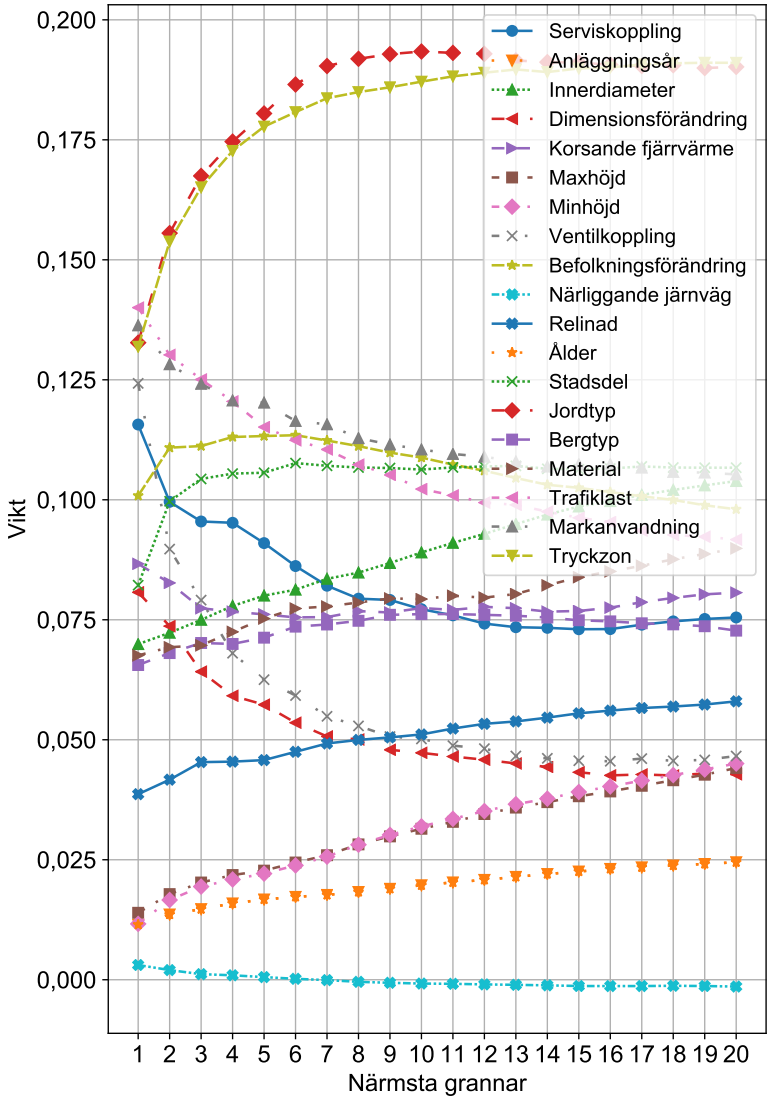


Figur 14. Den vänstra och högra figuren visar noggrannhet respektive förlust för modellen då alla attribut inkluderades.

4.1 ReliefF

ReliefF-algoritmen kördes med alla attribut, och antalet närmsta grannar, k som beaktades varierades i spannet 1 – 20 för att utvärdera när algoritmen presterade bäst. Figur 15 visar en graf över hur vikterna varierade med antalet närmsta grannar. Nominaldata är omvandlat till dummy-variabler, och i figuren visas den maximala vikten per attribut i respektive körning. Rangordningen av attributen baserade på respektive attributs vikt varierade mycket i början med antalet k. Därför visas både

$k = 5$ och $k = 10$ i tabell 9. Låg rang innebär att attributet är viktigt. Eftersom ordning och storlek på vikterna är relativt lika mellan $k = 5$ och $k = 10$ (tabell 9), baserades analysen på vikterna för $k = 10$ eftersom $k = 10$ rekommenderades i Kononenko m. fl. (1997). Figur C1 visar vilka värden vikterna går mot vid en väsentlig ökning av k. Det är främst resultatet av $k = 10$ som användes, men rangordningen då k är stort ($k = 1000$ i det här fallet) är också av intresse då det blir ett annat sorts mått på vikt än när k är litet (se avsnitt 5.2).



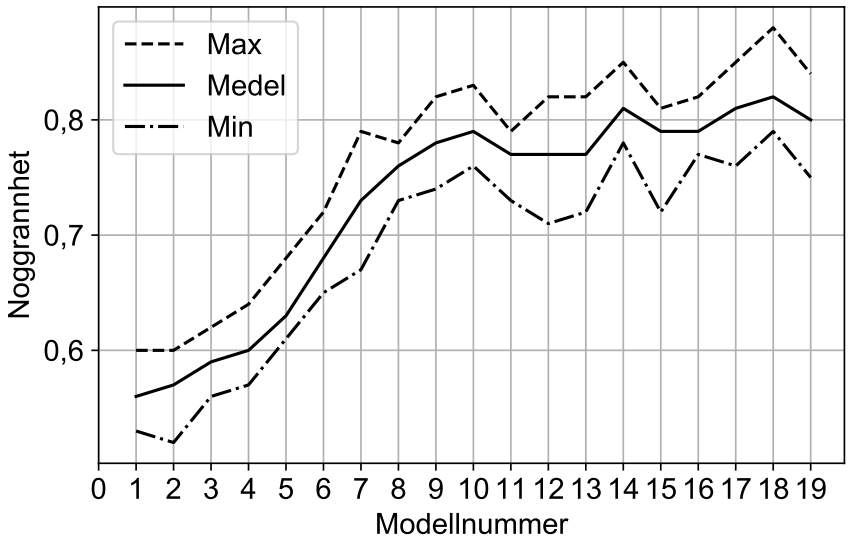
Figur 15. Attributens vikter från ReliefF-algoritmen.

Tabell 9. Vikter från ReliefF-algoritmen för k=10 och k=5. Låg rang indikerar att attributet är viktigt.

	Rank k=10	Rank k=5
Jordtyp	1	1
Tryckzon	2	2
Markanvandning	3	3
Befolkningsförändring	4	5
Stadsdel	5	6
Trafiklast	6	4
Innerdiameter	7	8
Material	8	10
Korsande fjärrvärme	9	9
Serviskoppling	10	7
Bergtyp	11	11
Relinad	12	14
Ventilkoppling	13	12
Dimensionsförändring	14	13
Minhöjd	15	16
Maxhöjd	16	15
Ålder	17	18
Anläggningsår	18	17
Närliggande järnväg	19	19

Hur modellens noggrannhet påverkas när fler och fler attribut läggs till i den ordning som ReliefF-algoritmen indikerar illustreras i figur 16. Modellnumret på figurens x-axel motsvar modellnumren i tabell C1. Notera att y-axeln startar på 0,50. Modellen är körd med tiofaldig korsutvärdering vilket innebär att modellen totalt kördes tio gånger men med olika tränings- och valideringsset för varje ny körning av modellen. Minimi-, medel- och maxvärde är tagna från dessa tio körningar. Som kan ses är det stor skillnad mellan modellens lägsta och högsta noggrannhet för respektive ny körning av modellen. Från grafen kan två olika beteenden skönjas: först ökar noggrannheten snabbt, för att sedan öka långsammare. Där lutningen är mindre är dessutom ökningen mindre stabil – den övergripande utvecklingen är att noggrannheten ökar, men för vissa steg minskar noggrannheten. Brytpunkten där noggrannheten

börjar öka långsammare sker vid modell 10, och de attribut som var med i modell 10 är redovisas i tabell 10.



Figur 16. Hur modellens noggrannhet förändras med antalet attribut. Vilka attribut som är med i respektive modell redovisas i tabell C1.

Tabell 10. De tio högst rangordnade attributen i ReliefF-algoritmen.

Attribut
Jordtyp
Tryckzon
Markanvändning
Befolkningsförändring
Stadsdel
Trafiklast
Innerdiameter
Material
Korsande fjärrvärme
Serviskoppling

Modell 63 i tabell 13 redovisar medelnoggrannheten då de attribut som inte förbättrar modellen i figur 16 exkluderats (Bergart, Relinad,

Ventilkoppling, Min- och Maxhöjd och Järnväg). Detta har ingen tydlig effekt på noggrannheten.

4.2 Wrapper

Resultatet av de två wrapperanalyserna (RFE baserad på MLR och på RFC) redovisas i tabell 11. De attribut med lägst nummer är de attribut som algoritmen funnit viktigast för att beskriva antalet läckor. Anledningen till att det kan förekomma stora hopp mellan nivåer i rangordningen är för att modellen kördes med kategoriska variabler som dummy-variabler och då får varje dummy-variabel en rang. Dummy-variablerna har sedan slagits ihop till sina respektive ursprungsvariabler och lägsta rangen (där en låg rang innebär att attributet är viktigt) behållits. Rangen för närliggande järnväg i RFC-RFECV, 26, innebär alltså att järnväg var det 26:e viktigaste attributet när alla dummy-variabler beaktades. Denna presentation har behållits fastän dummy-variablerna slagits ihop då det ger en tydligare inblick i attributets vikt. MLR-RFECV lyckades skapa en tydlig rangordning från viktiga till mindre viktiga attribut. RFC-RFECV lyckades endast särskilja tre attribut som mindre viktiga och därför användes inte resultatet från den analysen vid attributurvalet – resultatet vittnar dock om svårigheten i att särskilja attribut.

Tabell 11. Resultatet från RFE baserad på multinomial logistisk regression (MLR-RFECV) och random forest classificaiton (RFC-RFECV). Ju lägre siffra, desto viktigare är attributet.

Attribut	MLR-RFECV	RFC-RFECV
Anläggningsår	1	1
Minhöjd	1	1
Ålder	1	1
Innerdiameter	1	1
Material	2	1
Relinad	3	13
Markanvändning	4	1
Stadsdel	5	1
Korsande fjärrvärme	8	1
Trafiklast	9	1
Serviskoppling	12	1
Närliggande järnväg	16	26
Maxhöjd	20	1
Jordtyp	24	1
Bergtyp	25	1
Dimensionsförändring	37	1
Ventilkoppling	39	1
Befolkningsförändring	47	7
Tryckzon	48	1

Wrappermetoden MLR-RFE gav en tydligare rangordning av viktiga attribut än RFE med RFC. Eftersom MLR-RFE inte applicerades på ANN-modellen innebär det dock inte direkt att dessa attribut är viktiga. I tabell 12 jämförs skillnaden i noggrannhet mellan ReliefF och MLR-RFECV då 4 respektive 7 attribut inkluderades i respektive modell. För MLR-RFECV är det de sju första attributen då Relinad ignoreras. Relinad exkluderades eftersom endast ett fåtal ledningar hade det attributet; dessutom blev inte alla relinade ledningar i praktiken klassade som detta på grund av inkonsekvent inrapportering av relining.

Tabell 12. Jämförelse mellan ReliefF och MLR-RFECV.

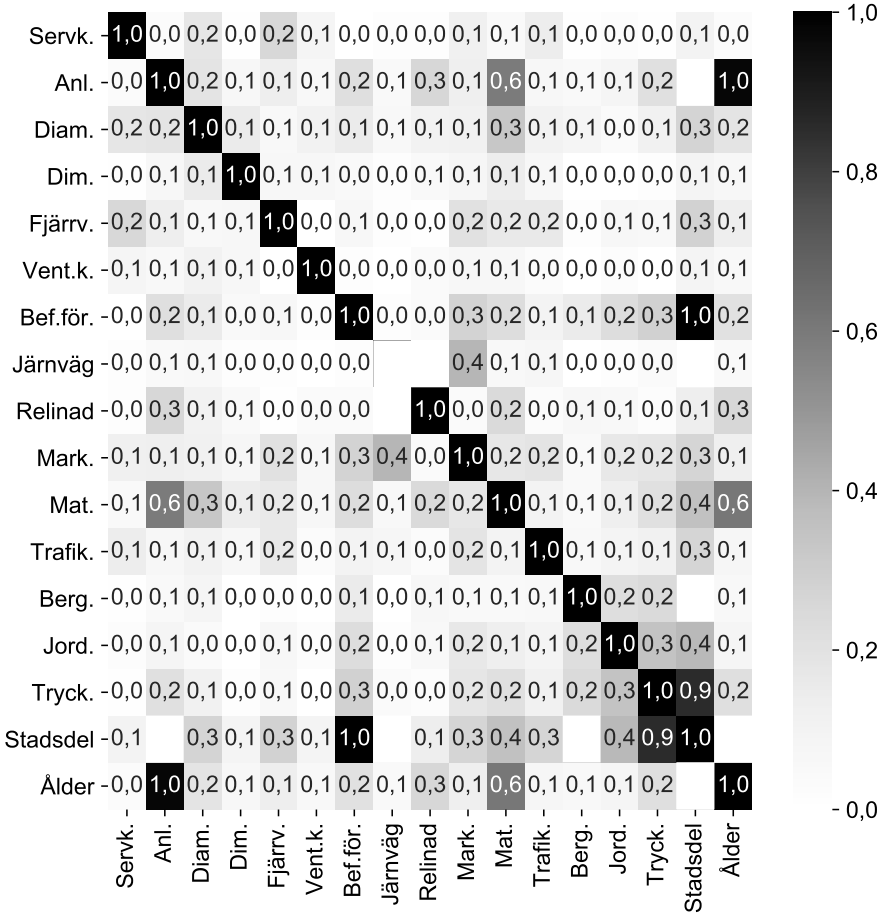
	Metod	
Antal attribut	ReliefF	MLR-RFECV
4	0,60	0,67
7	0,73	0,74

4.3 Korrelation

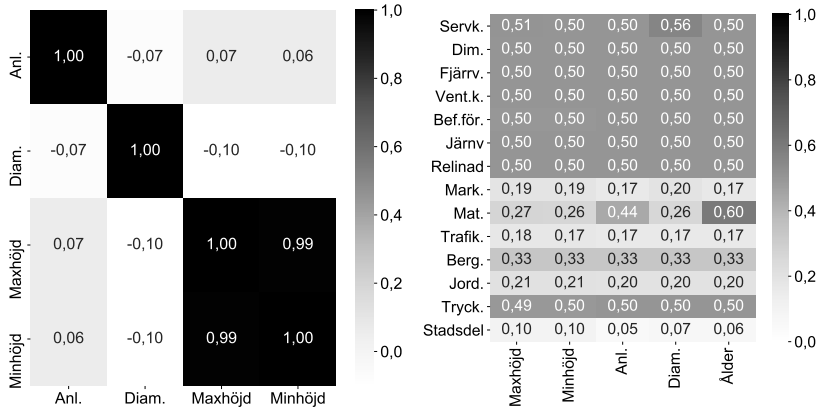
I figur 17 visas korrelation mellan kategoriska attribut. De vita rutorna utan angiven korrelation är de kombinationer av attribut som det inte gick att undersöka korrelation med hjälp av Cramérs V. Kombinationerna Anläggningsår och Ålder, och Befolkningsförändring och Stadsdel erhåller korrelationsvärden på 1. Detta beror på att Ålder är direkt beroende av Anläggningsår, och för den senare är orsaken att alla ledningsobjekt i en stadsdel har samma befolkningsförändring. Utöver dessa erhåller kombinationen Anläggningsår eller Ålder och Material ett värde på 0,6, och kombinationen Stadsdel och Tryckzon har ett korrelationsvärde på 0,9, vilket är relativt höga värden.

Figur 18 visar korrelationen mellan kontinuerliga attribut. Som kan ses är det endast höjderna som har betydande korrelation mellan varandra.

I figur 18 visas också korrelationen för kombinationen kontinuerliga och kategoriska attribut. Det är endast material och ålder, med ett korrelationsvärde på 0,6, som får ett korrelationsvärde noterbart över 0,5.

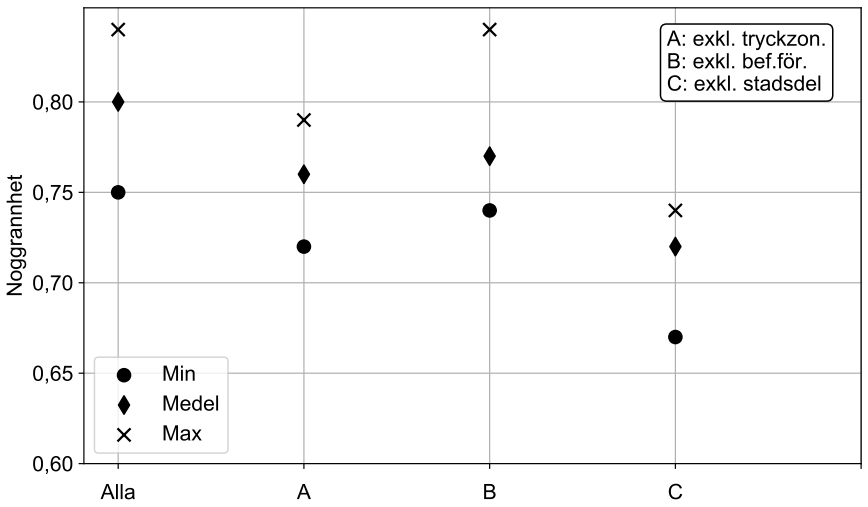


Figur 17. Värmekarta för korrelation, där korrelationen beräknats med Cramérs V. Ju mörkare nyans desto starkare korrelation. De vita rutorna utan korrelationsvärde är de kombinationer av attribut som som inte gick att undersöka med Cramérs V.



Figur 18. Den vänstra värmekartan visar korrelationen mellan de olika kontinuerliga attributen. Den högra bilden visar korrelationen mellan de kontinuerliga och de kategoriska attributen.

Resultatet från korrelationsanalysen påvisar tydlig korrelation inom grupperna Befolkningsförändring och Stadsdel, och Tryckzon och Stadsdel. Detta innebär inte att ett attribut inom respektive kombination bör uteslutas. Trots att korrelationen är hög kan samtliga attribut vara viktig för att förutsäga läcka på ledningsnätet. Figur 19 visar hur noggrannheten förändras då Tryckzon, Befolkningsförändring respektive Stadsdel utesluts från de 10 första attributen enligt ReliefF (tabell 9). I figuren påvisas att noggrannheten påverkas negativt när endera attribut tas bort.



Figur 19. Effekten av att ta bort de starkast korrelerande attributen från de 10 viktigaste attributen enligt ReliefF.

Utöver att korrelation skulle kunna användas för att utesluta ett attribut eller ta reda på ett attribut med hjälp av ett annat, påverkar korrelation även hur de andra utvärderingsmetoderna rangordnar attribut. RFE-RFECV kan välja bort ett attribut med stark korrelation medan ReliefF inte tar hänsyn till korrelation. Hur korrelation påverkar MLR-RFECV är inte känt i den här rapporten. Om MLR-RFECV likt RFC-RFECV tar hänsyn till korrelation skulle det kunna förklara varför exempelvis Tryckzon, som korrelerar starkt med Stadsdel, klassas som väldigt oviktig i MLR-RFECV men klassas som viktig i ReliefF.

4.4 Trial-and-error

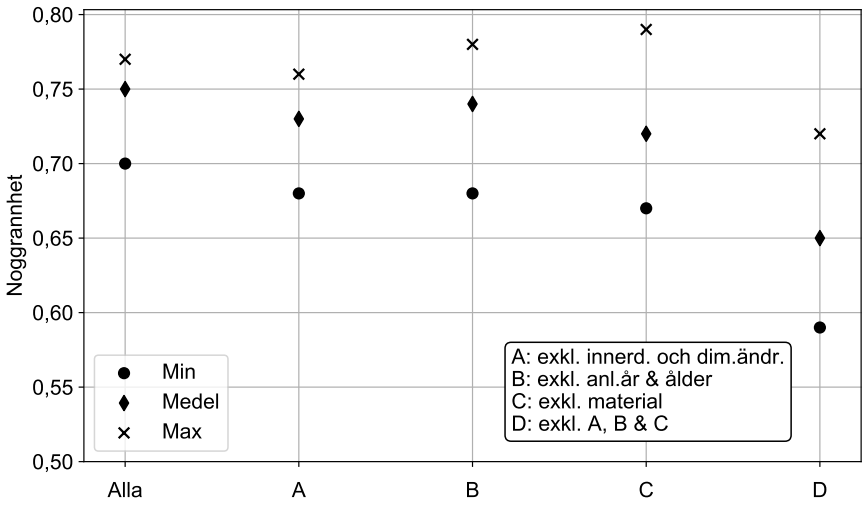
Utöver etablerade attributurvalsmetoder utvärderades attribut genom att först undersöka ANN-modellens prestation då endast ett attribut inkluderades. Till de modeller som presterade väl med endast ett attribut lades därefter till ytterligare ett attribut och så vidare. Kombination Anläggningsår, Ålder, Material och Trafiklast genererade en medelnoggrannhet på 0,72 (modell 36 i tabell 13).

4.5 Förenkling och generalisering av modellen

De fyra attributen Anläggningsår, Ålder, Material och Trafiklast gav en medelnoggrannhet på 0,72 (modell 36 i tabell 13) och dessa ingår i de enkla och generella attributen (tabell 8).

Modell 57 (tabell 13) är körd med de enkla och generella attribut som anges i tabell 8 och som ingår i de tio viktigaste attributen enligt ReliefF-algoritmen (det vill säga där brytpunkten i förbättringshastighet i figur 16 är). Detta gav en medelnoggrannhet på 0,68. Att modell 36 (tabell 13) presterar bättre trots färre attribut visar att även attribut som inte ingår i ReliefF-algoritmens topp tio attribut (figur 16) kan vara viktiga för modellen.

Modell 58 innehåller alla attribut i tabell 8 utan hänsyn till ReliefF-algoritmen. Denna modell erhöll en noggrannhet på 0,75 (tabell 13). Som nämndes i avsnitt 3.1.1 skulle anläggningsår kunna uppskattas genom att ta reda på när området byggdes ut, och material eventuellt utifrån anläggningsår. Korrelationsanalysen för Umeå visade dock att korrelationen mellan anläggningsår och stadsdel inte var särskilt hög (figur 18) och inte heller korrelationen mellan ålder och stadsdel var hög. Om information saknas om anläggningsår och material finns också risken att information om innerdiameter saknas. Modell 59 – 61 (tabell 13) undersöker därför effekten av att utesluta Material, Innerdiameter och Anläggningsår och Ålder. Dimensionsändring exkluderades tillsammans med innerdiameter eftersom ledningarnas innerdiameter används för att identifiera dimensionsändringar. Det attribut som påverkar noggrannheten mest är material, där noggrannheten faller från 0,75 till 0,72. Om inget av de fyra attributen tas med faller noggrannheten till 0,65 (modell 62). Modell 58 och uteslutandet av attribut illustreras också i figur 20.



Figur 20. Illustration av resultatet från modell 58-62 i tabell 13.

Tabell 13. Ett urval av körningar av ANN-modellen med olika attribut.

Modell	Min.	Med.	Max.	Attribut
36	0,68	0,72	0,76	Anläggningsår, Ålder, Material, Trafiklast
57	0,65	0,68	0,81	Material, Trafiklast, Serviskoppling, Innerdiameter, Jordart, Markanvändning.
58	0,7	0,75	0,77	Material, Dimensionsförändring, Trafiklast, Närliggande järnväg, Bergtyp, Serviskoppling, Ventilkoppling, Innerdiameter, Maxhöjd, Minimihöjd, Anläggningsår, Ålder, Jordtyp, Markanvändning.

Fortsätter på nästa sida

Tabell 13 – *Fortsättning*

Modell	Min.	Med.	Max.	Attribut
59	0,68	0,73	0,76	Material, Trafiklast, Närliggande järnväg, Bergtyp, Serviskoppling, Ventilkoppling, Maxhöjd, Minimihöjd, Anläggningsår, Ålder, Jordtyp, Markanvändning.
60	0,68	0,74	0,78	Material, Dimensionsförändring, Trafiklast, Närliggande järnväg, Bergtyp, Serviskoppling, Ventilkoppling, Innerdiameter, Maxhöjd, Minimihöjd, Jordtyp, Markanvändning.
61	0,67	0,72	0,79	Dimensionsförändring, Trafiklast, Närliggande järnväg, Bergtyp, Serviskoppling, Ventilkoppling, Innerdiameter, Maxhöjd, Minimihöjd, Anläggningsår, Ålder, Jordtyp, Markanvändning.
62	0,59	0,65	0,72	Dimensionsförändring, Trafiklast, Närliggande Järnväg, Bergtyp, Serviskoppling, Ventilkoppling, Maxhöjd, Minimihöjd, Jordtyp, Markanvändning.

Fortsätter på nästa sida

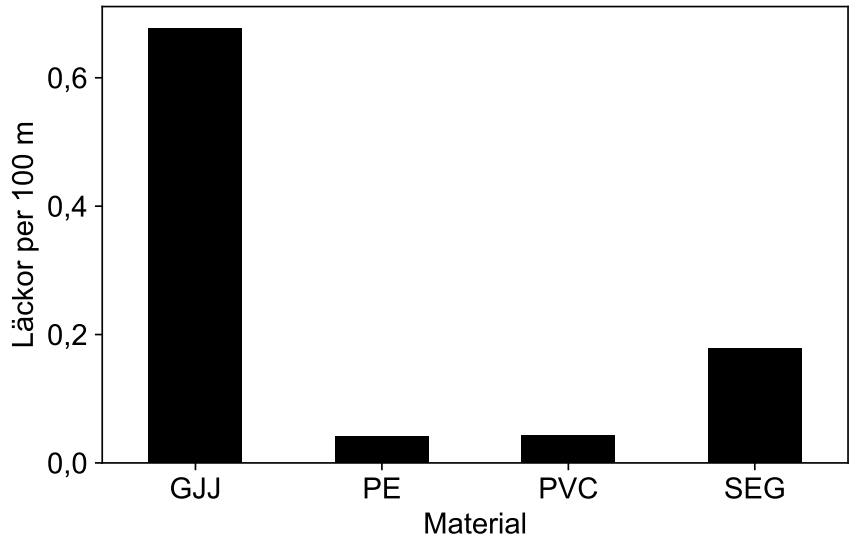
Tabell 13 – *Fortsättning*

Modell	Min.	Med.	Max.	Attribut
63	0,78	0,81	0,85	Jordart, Tryckzon, Markanvändning, Befolkningsförändring, Stadsdel, Trafiklast, Innerdiameter, Material, Korsande fjärrvärme, Serviskoppling, Dimensionsändring, Anläggningsår, Ålder

4.6 Attributs påverkansgrad på ledningsnätet

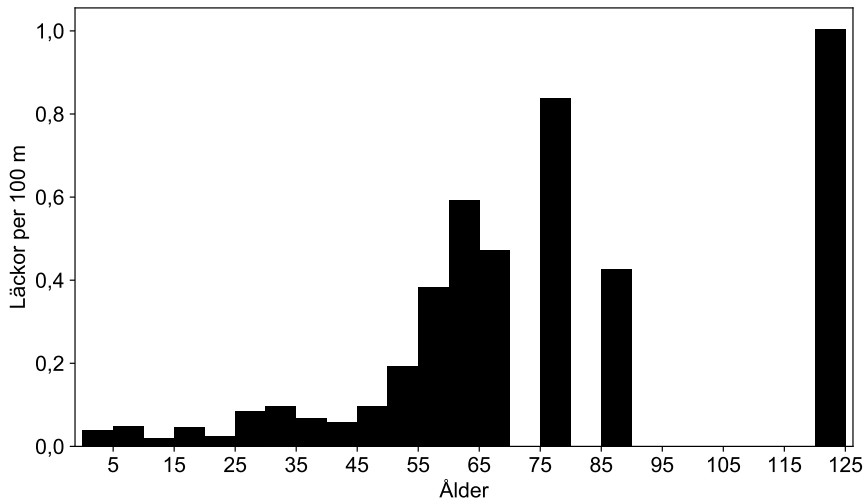
Denna analys baserar sig på en körning av ANN-modellen då alla attribut inkluderades. Det är den körning som redovisas i figur 14 och det är inte samma körning som modell 19 i tabell C1. För de grafer där uppdelnings gjorts på material förekommer ibland siffror över staplar. Dessa siffror markerar hur många värden stapeln baseras på. Ifall stapeln baseras på 20 värden eller fler visas ingen siffra. Figur 29 som visar Umeås stadsdelar är ett undantag. Då visar siffrorna respektive stadsdels ålder. Gränsen för läcka har satts till en prediktion över 0,75. Gränsen för läcka diskuteras ytterligare i avsnitt 5.1.

Material visade sig vara ett viktigt attribut då attributens vikt analyserades. Figur 21 visar antalet läckor per hundra meter för de olika materialen. Som kan ses är gjutjärn klart överrepresenterat medan både PE och PVC har en låg läckfrekvens.



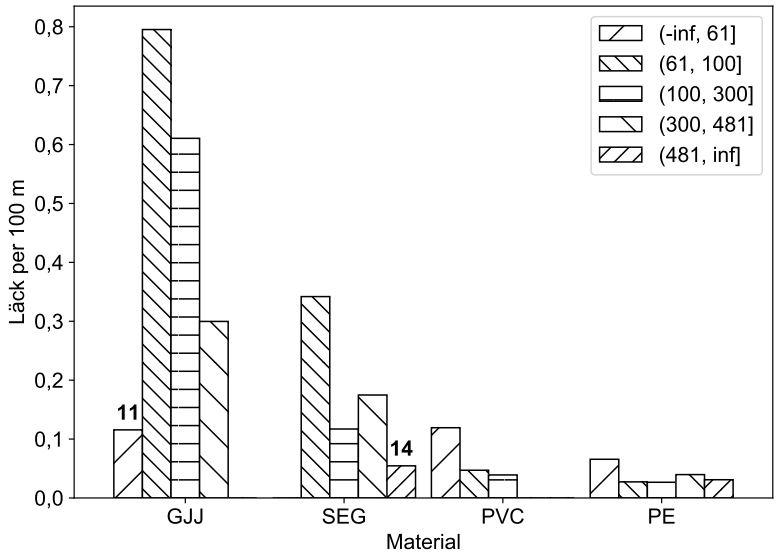
Figur 21. Läckfrekvens för de olika materialen.

Figur 22 visar läckfrekvensen baserat på ålder. Frekvensen verkar öka med ålder, men är relativt konstant fram till 45 år. Efter en ålder på 45 ökar läckfrekvensen markant.



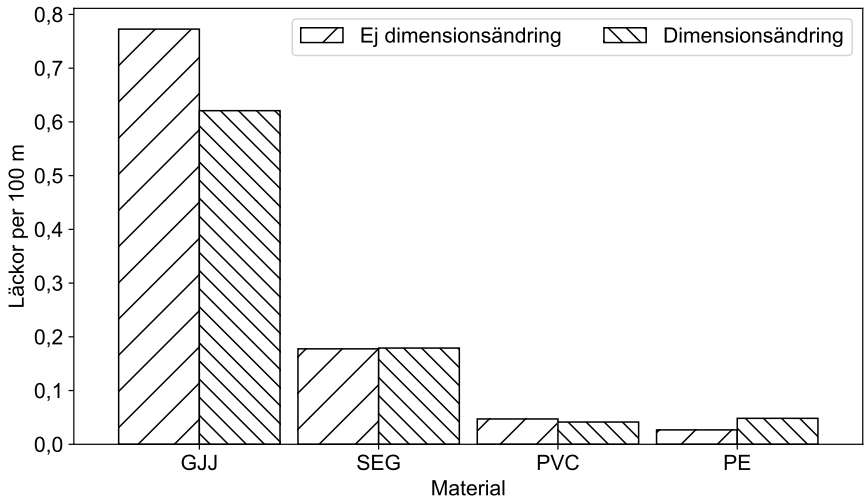
Figur 22. Läckfrekvens baserat på ålder.

Ledningar med liten innerdiametrar har i tidigare forskning indikerats vara överrepresenterade för läckor (Sundahl, 1996). Figur 23 visar läckfrekvens för grupper av diameter och rörmaterial. För alla material minskar läckfrekvensen överlag med ökad diameter, men flera undantag förekommer. För PE ökar läckfrekvensen för de två största diametergrupperna. Gjutjärn sticker ut där läckfrekvensen är mycket lägre för den minsta gruppen jämfört med de tre efterföljande grupperna.



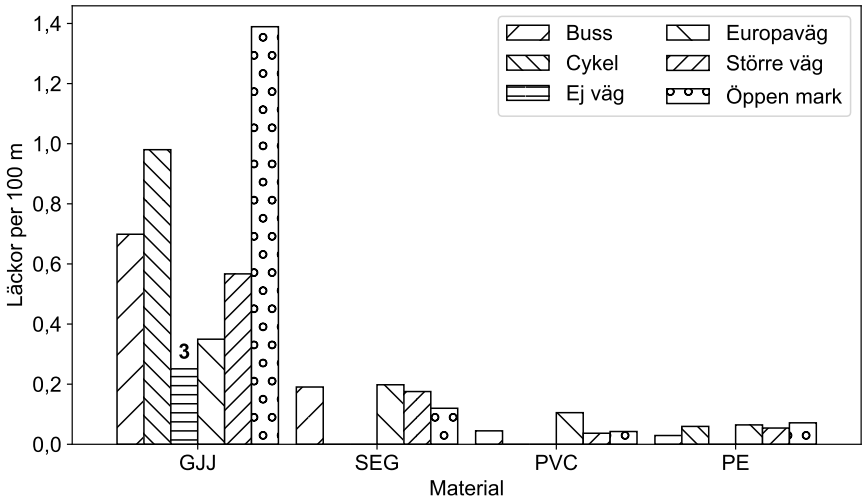
Figur 23. Läckfrekvens för diameter. Måtten är i millimeter.

Enligt Malm, Horstmark, Larsson m. fl. (2011) har övergång från en större ledningsdimension till en mindre varit ett problem för PE-ledningar, och ANN-modellen indikerar att så är fallet även för Umeå (figur 24). Tilläggas bör att modellen endast tar hänsyn till dimensionsändring, men eftersom strömningsriktningen kan variera i ett cirkulationsnät är det svårt att säga om vattnet går från större ledning till mindre ledning eller vice versa. För de andra materialen har *Ej dimensionsändring* lika hög eller högre läckfrekvens.



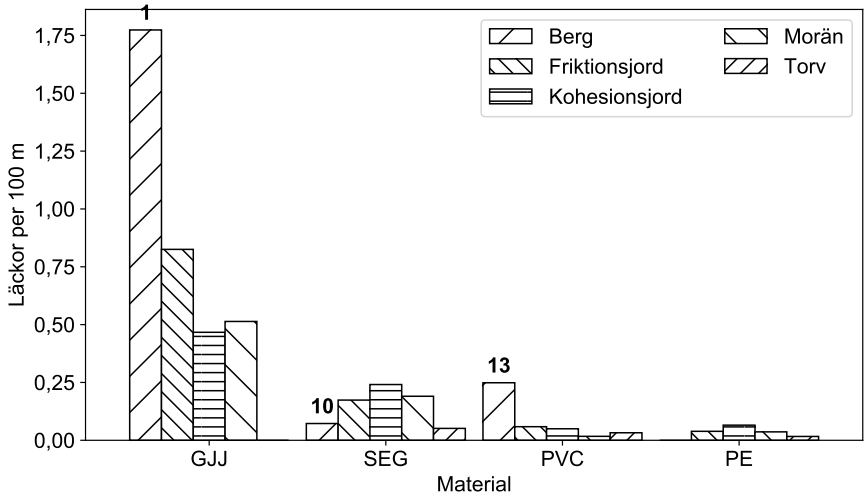
Figur 24. Läckfrekvens för dimensionsändring grupperat på material.

För både gjutjärn och PVC har en tidigare analys visat att trafiklaster är en viktig faktor för läckage (Sægrov, 1998). Figur 25 visar hur läckor fördelas över de olika trafiklasterna. Både cykelväg och öppen mark är klart överrepresenterade för gjutjärn. De tyngre lasterna, buss-, europa- och större väg har klart lägre frekvens än öppen mark och cykelväg. För segjärn är buss-, europa-, och större väg överrepresenterat. För PVC har europaväg högst läckfrekvens. För PE är läckfekvensen jämnt fördelad mellan material, men buss och ej väg är underrepresenterade..



Figur 25. Läckfrekvens för olika trafiklaster grupperat på ledningsmaterial.

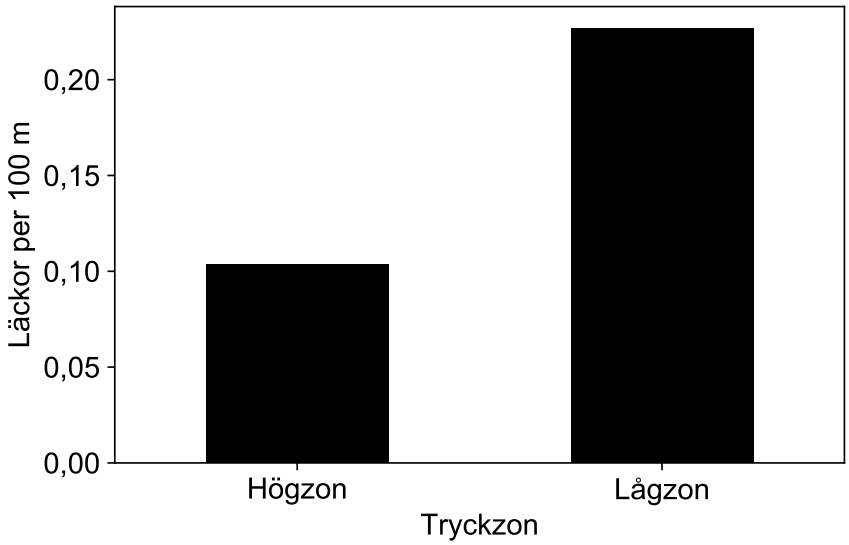
Jordmaterial kan påverka risk för sättningar, balkverkan och korrosion samt påverka ledningsbädden. Speciellt jordlast, sättningar och balkverkan har varit viktiga faktorer för läckage på gjutjärn i tidigare studier (Malm, Horstmark, Larsson m. fl., 2011; Sægrov, 1998), och undermålig ledningsbädd har varit särskilt viktigt för PVC (Sægrov, 1998). Korrosion är något som observerats vara viktigt för segjärn med avseende på läckage (Sægrov, 1998). Figur 26 visar läckfrekvens för jordmaterial grupperat på material. Berg är klart överrepresenterat för gjutjärn och PVC gällande läckfrekvens, medan friktionsjord, kohesionsjord och morän är överrepresenterade för segjärn. För PE är läckfrekvensen högst för kohesionsjord.



Figur 26. Läckfrekvens för jordmaterial grupperat på ledningsmaterial.

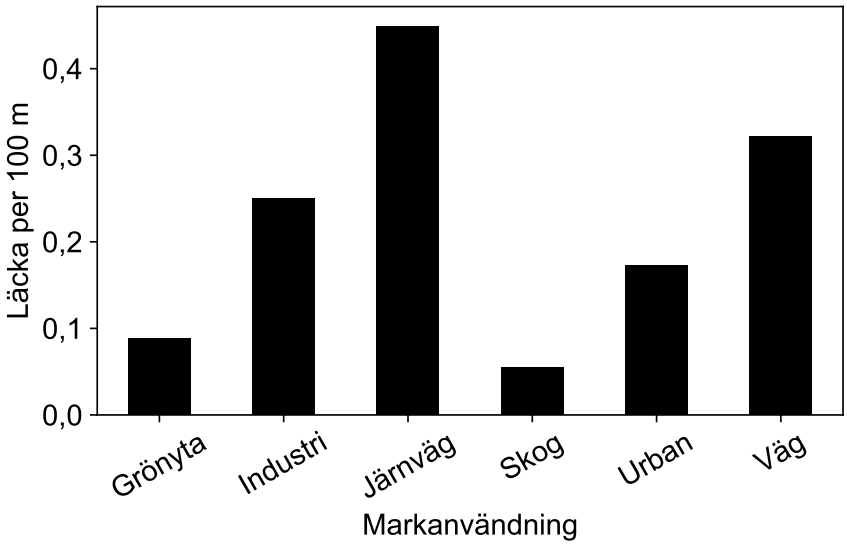
Utöver faktorer som visat sig vara viktiga i tidigare undersökningar kan det vara intressant att undersöka attribut som visat sig vara viktiga för att uppnå gott resultat i den ANN-modell som använts i det här arbetet. Många av de viktigaste attributen från ReliefF-algoritmen (tabell 9) har tagits upp ovan med hänsyn till tidigare forskning. Tryckzon, Befokningsförändring, Stadsdel, Korsande fjärrvärme och Serviskoppling har inte berörts.

I figur 27 redovisas läckfrekvenserna för Umeås två tryckzoner. Läckfrekvensen för lågzonen är mer än dubbelt så hög som för högzonen.



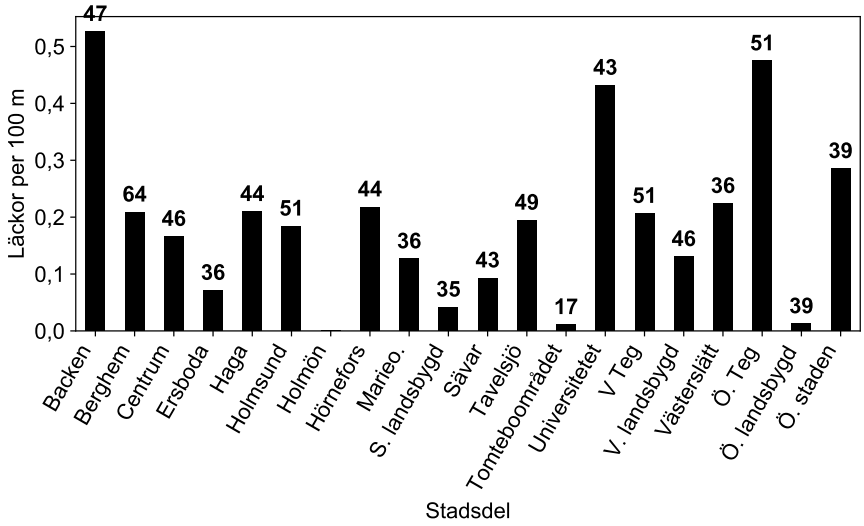
Figur 27. Läckfrekvens för de två olika tryckzonerna.

Läckfrekvensen för respektive markanvändning redovisas i figur 28. De markanvändningar som kan antas vara förknippade med hög last: väg, urban, järnväg och industri är tydligt överrepresenterade.



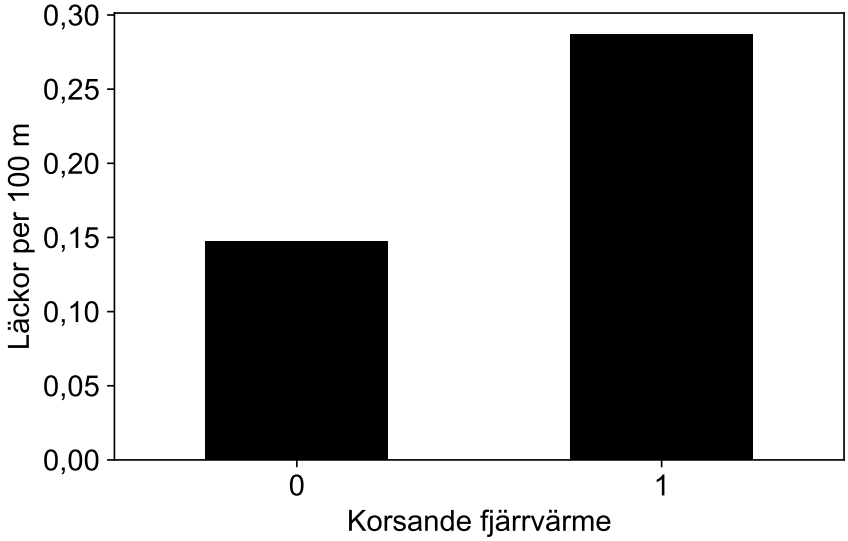
Figur 28. Läckfrekvens för markanvändning.

Läckfrekvensen för olika stadsdelar redovisas i figur 29. Läckfrekvensen skiljer sig mycket mellan olika stadsdelar. Ovanför varje stapel visas medelåldern för ledningsnätet i respektive stadsdel.



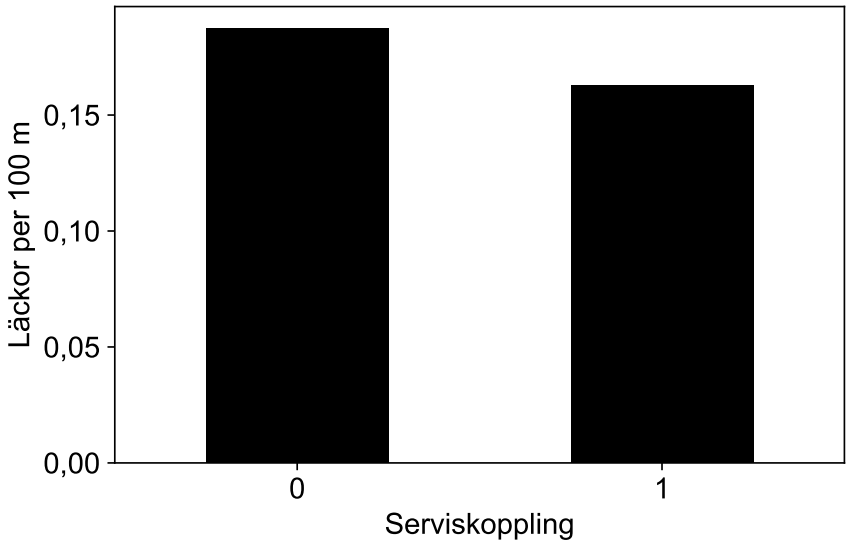
Figur 29. Läckfrekvens för de olika stadsdelarna. Siffran över respektive stapel är medelåldern för respektive stadsdels ledningsnät.

Läckfrekvensen för korsande fjärrvärme redovisas i figur 30. Läckfrekvensen för korsande fjärrvärme är nästan dubbelt så hög som utan korsande fjärrvärme.



Figur 30. Läckfrekvens för ej korsande fjärrvärme, 0, och korsande fjärrvärme, 1.

I figur 31 visas läckfrekvensen för serviskoppling. Läckfrekvensen är något högre för ledningar utan serviskoppling.



Figur 31. Läckfrekvens för ej serviskoppling, 0, och serviskoppling, 1.

5 Diskussion

I det här avsnittet kommer först modellens prestation utvärderas, varefter attributurvalsmetoderna och deras resultat diskuteras. Sedan avhandlas avvägningar som gjorts för både indata och inom tränings- och utvärderingsskedet. Därefter diskuteras resultatet kring enkla och generella attribut samt ANN-modellens roll i underhållsplanering. Till sist avhandlas läckfrekvens kopplat till olika attribut baserat på modellens prediktioner.

5.1 Utvärdering av modell

Ett återkommande problem i studien var osäkerheten i facit, beskrivet i avsnitt 3.3. Detta innebär att den optimala noggrannheten är okänd: en noggrannhet på 100 procent skulle innebära att det inte finns några oidentifierade läckor vilket inte är realistiskt. Denna problematik är främst ett problem när resultatet från ANN-modellen används för att utvärdera läckfrekvensen på Umeås ledningsnät. Om det antas att modellen identifierat samband – om än inte tillräckligt tydliga för att klassificera ledningar med hög säkerhet – är det främst relationen mellan noggrannheten för alla attribut och ett begränsat antal attribut som är viktig när attribut utvärderas. Som referensvärde för noggrannhet då alla attribut inkluderades användes medelnoggrannheten från modell 19 (tabell C1 och figur 16): 0,80. Både modell 17 och 18 presterade dock bättre med medelnoggrannheter på 0,81 respektive 0,82. Framgent kommer ordet *läckledning* användas för att beskriva en ledning som haft en historisk läcka.

Baserat på den läckstatistik som redovisas i tabell 1 i avsnitt 1.1 har Umeå haft ett snitt på 80 läckor per år de senaste åtta åren, 2010 och 2011 uteslutna då de värdena visar stor diskrepans med statistik inrapporterat till VASS. Att endast utvärdera modellen mot inrapporterade läckor är inte ett bra mått då det inte inkluderar oidentifierade läckor: det kan finnas många små läckor som inte har identifierats. Eftersom det till viss del är samma faktorer som leder till små läckor som till stora bör ANN-modellen identifiera även dessa läckor. Läckstatistik kan dock

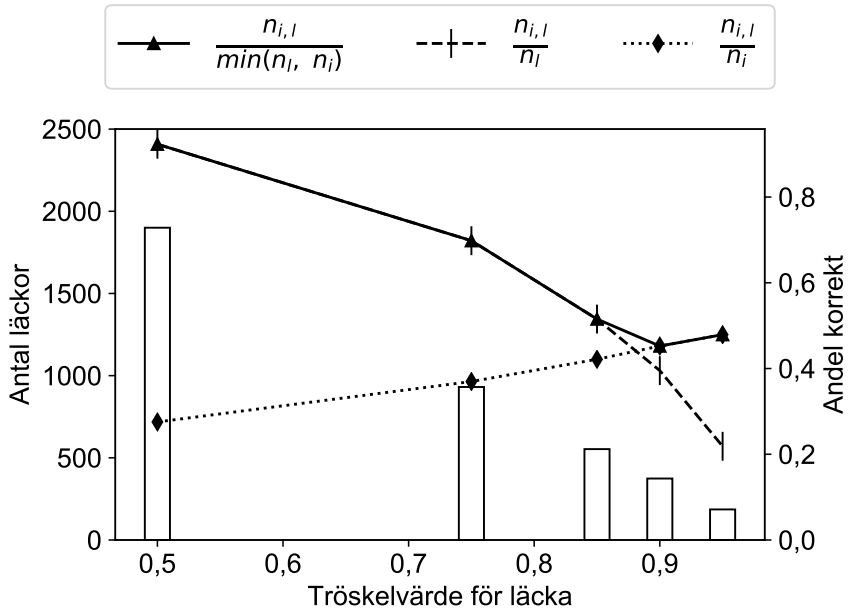
ge en indikation på vilken storleksordning på läckfrekvens som kan förväntas.

En aspekt som komplicerar utvärderingen är att utdata från ANN-modellen är en prediktion mellan noll och ett. Ju högre prediktion desto säkrare är modellen på att en läcka förekommer, och vilken nivå tröskeln för läcka sätts till påverkar hur många ledningar som klassas som läcka.

Figur 32 visar hur antalet läckor varierar beroende på tröskelvärdet för läcka. Värdena är exklusive redan identifierade läckor. Dessa läckor har lagats, men ledningarna kan ändå ha en förhöjd risk för läcka och därför är det inte självklart att alla dessa ledningar ska uteslutas.

Figur 32 visar också andelen läckledningar som modellen klassificerat korrekt. Andelen korrekt klassade enligt den streckade linjen minskar konstant. Det är för att nämnaren är det totala antalet läckledningar. Sätts tröskelvärdet för läcka till noll kommer modellen att klassa alla ledningar som läcka, inklusive läckledningar, och andelen korrekta hade då varit ett. Den prickade linjen har antalet identifierade ledningar som nämnare, och antalet identifierade ledningar minskar med ökat tröskelvärde. Att denna linje ökar konstant innebär att andelen läckledningar som identifieras ökar med tröskelvärdet. Den heldragna linjen är definierad som den streckade linjen då antalet identifierade ledningar är större än antalet läckledningar, och som den prickade linjen då antalet läckledningar är mindre än antalet läckledningar.

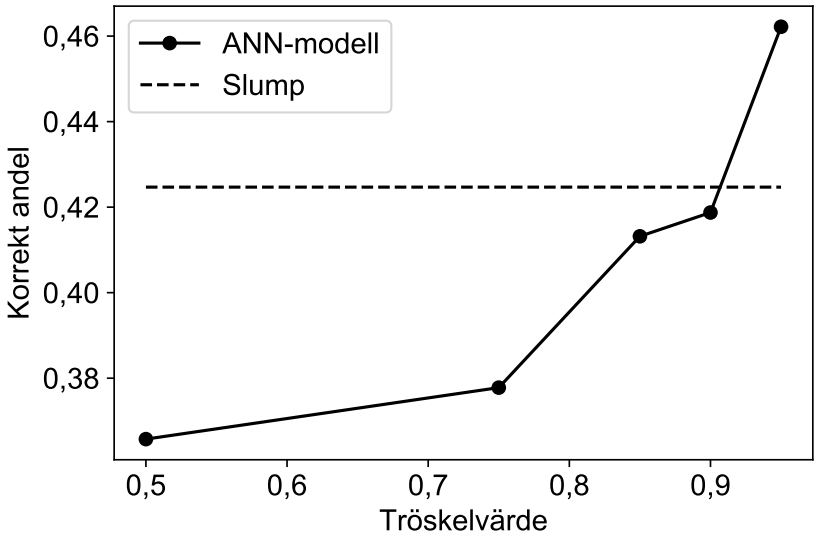
Modellens förlust mäter skillnaden mellan den prediktion en ledning ges och det korrekta värdet. Denna förlust har varit stor i alla modeller (ca 0,5) vilket visar att modellen haft svårt att lära sig identifiera ledningar med risk för läckor. Detta är troligtivs inte endast ett resultat av osäkerheten i facit. Ett annat problem är den begränsade datamängden vilket illustreras tydligt med att noggrannheten ökade ungefär med 10 procentenheter då läckledningarna duplicerades (avsnitt 3.3). Utöver ökad indata är det möjligt att skapa en mer komplicerad modell, men detta skulle öka risken för överpassning.



Figur 32. Figuren visar antalet läckor modellen identifierar, samt olika mått på hur många ledningar modellen klassar korrekt. $n_{i,l}$ är antalet identifierade ledningar som också är läckledningar, n_l är antalet läckledningar och n_i är antalet identifierade ledningar.

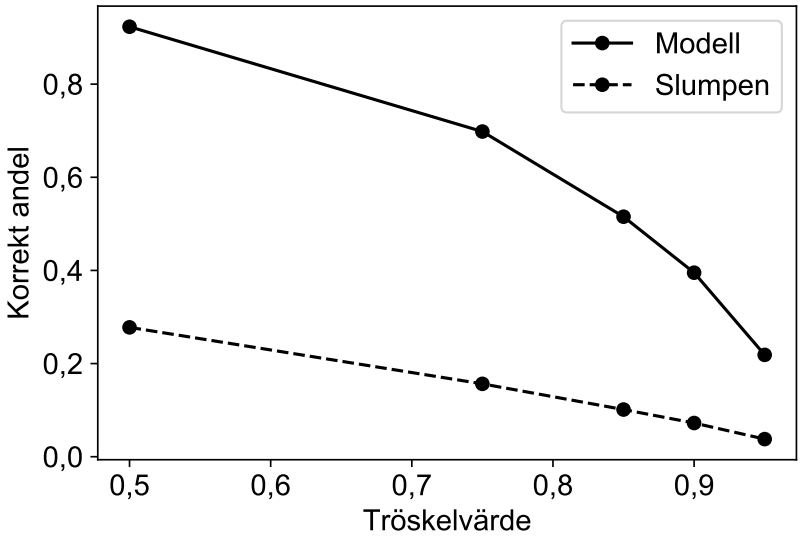
Vakin har gjort en utvärdering av ledningsnätet där både sannolikhet för läcka och konsekvens av läcka beaktas. Utifrån detta klassas respektive ledningsobjekt som antingen *ingen risk*, *kostnadsnytta*, *håll koll* eller *högsta prio*. I gruppen håll koll ingår många objekt: 4 436 av 11 554 stycken. I den högsta klassen *högsta prio* ingår 239 objekt. Denna klassning kan jämföras med ANN-modellens klassning, även om de inte är helt jämförbara eftersom ANN-modellen inte inkluderar konsekvens. Figur 33 visar andelen ledningar modellen klassar som läcka och som ingår i håll koll eller högsta prio. Lite mindre än hälften av ledningarna som klassas som läcka ingår i någon av de två grupperna med högst risk då gränsen för läcka sätts till 0,95. Eftersom modellerna inte bygger på samma premisser innebär detta inte nödvändigtvis att ANN-modellen presterar dåligt. Väljs 357 ledningar slumpmässigt, varav 4 009 är korrekta (det vill säga riskklass 3 eller 4 enligt Vakins modell) av totalt

9 440 är den förväntade andelen korrekta 42 procent. ANN-modellen presterar alltså marginellt bättre än slumpen.

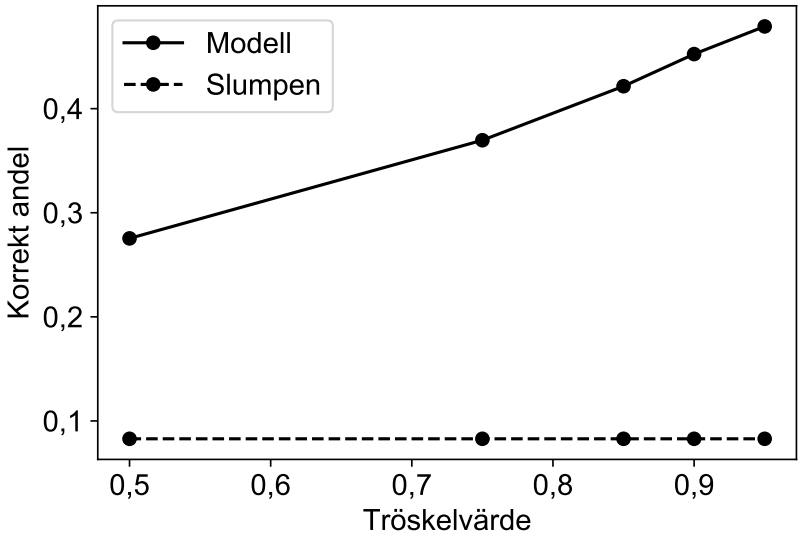


Figur 33. Andel ledningar som klassas som läcka enligt ANN-modellen och ingår i de två mest riskfyllda grupperna.

Ett annat mått på modellens prestation är hur många läckledningar modellen identifierar jämfört med slumpen. Detta redovisas i figur 34. Efter det tredje värdet väljs färre ledningar ut än vad som kan vara korrekt och då kan inte en andel på ett uppnås. Ett annat mått är därför andelen korrekt klassade ledningar. Detta redovisas i figur 35.



Figur 34. Andel korrekt klassade läckledningar där nämnaren är antalet historiska läckledningar.



Figur 35. Andel korrekt klassade läckledningar där nämnaren är antalet identifierade ledningar.

Ur flera aspekter hade en högre modellprestation varit önskvärd, men modellen presterar ändå betydligt bättre än slumpen. Utan att introducera mer data skulle modellen kunna förbättras genom att göra den mer komplicerad men detta ökar också risken för överpassning. I och med denna risk måste en utveckling av modellen göras på ett väl genomtänkt sätt så att modellens prestation kan utvärderas på ett bra sätt. Detta kan delvis göras i enlighet med vad som gjorts här, men ytterligare metoder kan vara spatiala analyser där exempelvis problemområden identifieras och jämförs med verkligheten.

5.2 Utvärderingsmetodernas prestation

Figur 16 visar hur noggrannheten för ANN-modellen förändras då attribut läggs till i den ordning ReliefF-algoritmen föreslår. Figuren visar tydligt att fler inparametrar överlag ger högre noggrannhet. Detta är som förväntat – ANN-modeller ska vara robusta mot brus. Detta är dock inte ett allenarådande tillstånd. Efter modell 10 i figuren kan en svag ökning av noggrannheten ses men med stor variation mellan körningar. Noggrannheten har varierat även när samma attribut använts och den variation som ses efter modell 10 kan bero på detta och inte att vissa attribut har negativ påverkan på modellen. Modell 63 i tabell 13 visar modellens noggrannhet utan de attribut som inte förbättrade modellens prestation (Bergart, Relinad, Ventilkoppling, Min- och Maxhöjd och Järnväg), och en noggrannhet på 0,81 uppnåddes då. Detta är ungefär samma noggrannhet som för modell 19: 0,8 (se figur 16). De fem attributen verkar alltså inte ha någon tydlig påverkan på modellens prestation.

Att ökningen är snabb och stabil för de tio första attributen i figur 16 påvisar att ReliefF-algoritmen gör ett bra jobb i att identifiera viktiga attribut. De efterkommande attributen har endast en begränsad påverkan på noggrannheten. Vilka attribut som ingår i modell 10 respektive 19 redovisas i tabell C1. De mindre viktiga attributen, attribut 11 – 19, redovisas i tabell 14.

Tabell 14. De tio viktigaste attributen enligt ReliefF-algoritmen. De är uppradade i enlighet med deras rang i ReliefF-algoritmen.

Attribut
Bergtyp
Relinad
Ventilkoppling
Dimensionsändring
Minimihöjd
Maxhöjd
Anläggningsår
Ålder
Närliggande järnväg

Att anläggningsår, ålder och minimihöjd är lågt rankade i ReliefF-algoritmen är märkligt då de har stor vikt i RFE-algoritmen baserade på MLR. Dessutom presterade en modell med attributen Anläggningsår, Ålder, Material och Trafiklast väl med hänsyn till antalet attribut (modell 36 i tabell 13). En förklaring kan vara att ReliefF-modellen underskattar kontinuerliga attribut ifall även kategoriska attribut förekommer. Ytterligare en förklaring till denna diskrepans relaterar till skillnaden mellan ett stort antal närmsta grannar, k, och ett fåtal k i ReliefF-algoritmen. Figur C1 visar resultatet av ReliefF-algoritmen när $k = 5, 10, 50, 100, 300, 500$ och 1000. Vid $k = 1000$ är Anläggningsår, Ålder, Material, Minimihöjd och Maximihöjd de viktigaste attributen. Detta kan tolkas som att det finns viktigare attribut för att förutspå läckage på en vattenledning än dessa, men attributen innehåller ändå information om risken för läckage. Ett sätt att tänka på det är att när k blir stort kommer den aktuella ledningssträckan att jämföras med väldigt många andra ledningssträckor som kommer att vara mer och mer olik ledningssträckan ju större k är. Många attribut blir då inte längre viktiga för utfallet eftersom eventuell viktig information drunknar i all oviktig information. För de attribut som ändå erhåller en hög vikt vid ett stort k verkar ett samband råda för ledningsnätet i sin helhet, och inte bara för liknande ledningar. Detta samband kan dock vara svagt och inte vara viktigt då k är litet och viktigare lokala samband återfinns. Detta kan indikera att Anläggningsår,

Ålder, Minimi- och maximihöjd trots låg vikt vid $k = 10$ kan vara bra att ha med i en ANN-modell.

Den ökade dimensionaliteten efter att nominala attribut omvandlats till dummy-variabler, beskrivet i avsnitt 3.2 och 3.4, med en ökning från 19 till 58 attribut kan ha försvårat för alla tre attributurvalsmetoder att identifiera viktiga attribut. Oförmågan var tydligast för RFE-RFECV, men även för ReliefF var det många vikter som låg nära varandra (figur 15). För RFC-RFECV finns metoder som hanterar kategoriska attribut på ett effektivt sätt, men någon sådan är ännu inte tillgänglig i Scikit-Learn. För ReliefF finns det ingen metod som jag känner till. ReliefF mäter avstånd mellan attribut i dimensionsrummet och det går inte att mäta avstånd mellan kategoriska attribut, vilket komplicerar hanterandet av kategoriska variabler. I stället för att skapa ett nytt attribut för varje dummy-variabel kan varje kategori tilldelas en siffra, men det skulle innebära att avståndet skiljer sig mellan olika kategoriska variabler vilket inte är sant.

Korrelation undersöktes på flera olika sätt för att identifiera ifall stark korrelation rådde mellan några attribut. Tanken var att ifall exempelvis Anläggningsår och Material hade stark korrelation hade det varit tillräckligt att ha med ett av de attributen. Att de korrelerar starkt innebär dock inte att det ena attributet kan uteslutas utan att modellen försämras (Guyon & Elisseeff, 2003), men den här studien gick ut på att minimera antalet attribut så en viss förlust i prestation hade varit acceptabel. De attribut som uppvisade stark korrelation var Max- och Minimihöjd, Anläggningsår och Ålder, Stadsdel och Tryckzon och Stadsdel och Befolkningsförändring. Att endast behålla ett av attributen i respektive kombination Max- och Minimihöjd och Anläggningsår och Ålder är inte befogat eftersom höjddata är tillgängligt via lantmäteriet och extraheras på samma sätt för både max- och minimihöjd, och ålder fås direkt från anläggningsår. Att utesluta endera av Stadsdel, Befolkningsförändring eller Tryckzon resulterade i lägre noggrannhet (figur 19) och om det är möjligt att erhålla alla dessa är det därför rekommenderat. Ett annat sätt att använda korrelation på är att uppskatta värdet på en parameter utifrån en annan. Det skulle exempelvis vara möjligt att bestämma Befolkningsförändring eller Tryckzon från Stadsdel med relativt hög

träffsäkerhet.

5.3 Träning av modellen

Vid tränandet av modellen behövdes flera överväganden göras. Dessa beskrivs i följande avsnitt.

En effekt av osäkerheten i facit är att det är svårt att bedöma ifall ökad noggrannhet är ett resultat av att modellen i praktiken presterar bättre, eller ifall den blir bättre på att klassa ledningar med oidentifierade läckor som nollor: ej läcka. Detta är särskilt aktuellt vid beslutet att dubblera ledningarna med läckor, där dubbleringen höjde noggrannheten med tio procentenheter. Att dubblera värden är inte ett normalt förfarande i en ANN-modell, utan det var ett nödvändigt ont. När modellen kördes utan att dubblera antalet läckor identifierade modellen endast fyra ledningar som läcka, och ingen av de historiska läckorna blev klassad som läcka. Ökningen av noggrannheten med tio procentenheter anses därvidlag ha förbättrat modellen även i praktiken.

Effekten av att exkludera testdata blev att den sista kontrollen mot överpassning försvann. Den främsta anledningen till att detta gjordes var behovet av mer data att träna modellen på. En variant av testdata kunde ha erhållits genom att skapa ett testdataset genom att använda samma ettor, men byta ut nollorna. Kontrollen av överpassning hade blivit sämre, men testprocessen skulle inte ha medfört att modellen fick mindre data att träna på. Denna utvärderingsmetod bör undersökas närmare för att undersöka hur väl den presterar, då det vore önskvärt att kunna inkludera testdata på något sätt. En utvärdering enligt förfarandet i avsnitt 5.1 kan vara bra, men den säger ingenting om hur modellens prestation påverkas då modellen får se ny data. Därför kan testdata och en noggrann utvärdering komplettera varandra.

Den ökade dimensionaliteten, beskriven för ReliefF och RFC ovan, blir även ett problem i ANN-modellen. Antalet noder står fritt att välja och ju fler noder som används, desto mer komplicerade samband kan modellen lära sig. Därför ökar också risken för att överpassa modellen med antalet noder. I det här arbetet har antalet noder valts att öka linjärt med antalet attribut inklusive dummy-variabler: ($n_{attribut}$ + 10). Att

inkludera Stadsdelar ökar antalet noder med 20, vilket är en markant ökning med tanke på att endast ett attribut läggs till. I vissa körningar av modellen märktes en tydlig ökning av noggrannheten då Stadsdel inkluderades och detta kan ha varit ett resultat av att antalet noder ökade. I efterhand kan det ha varit ett bättre alternativ att basera antalet noder på antalet unika attribut.

5.4 Förenkling och generalisering av modell

Utvärderingen av enkla och generella attribut visade att modellen kan uppnå en noggrannhet nära noggrannheten då alla attribut inkluderades, men att vissa attribut som kan vara svåra för kommuner att få fram är viktiga för att erhålla en god noggrannhet – detta gällde främst ledningsmaterial. Modell 58 (tabell 13) innehåller alla attribut som kommuner bör kunna få fram och som generaliserar väl. Dessa attribut redovisas i tabell 8. Noggrannheten för modellen med dessa var 0,75. Utan material blev noggrannheten 0,72, och utan material, innerdiameter, dimensionsändring, anläggningsår och ålder blev noggrannheten 0,65 (se figur 20). Detta indikerar att även om dessa attribut kan vara svåra att anskaffa, är de viktiga för att erhålla en högre noggrannhet.

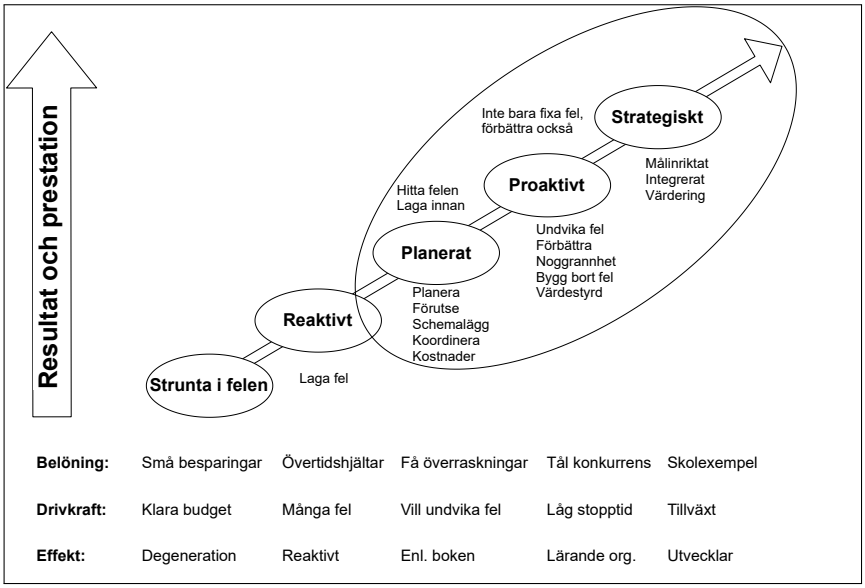
5.5 ANN för strategisk underhållsplanering

I den här rapporten har ledningsnätets kondition främst behandlats som läcka eller inte läcka. I förlängningen finns det ett värde av att gå ifrån att endast diskutera läckage, och i stället prata om kondition och prestation – detta för att kunna planera underhållet på ett effektivt sätt. Ju bättre koll på ledningsnätet en VA-organisation har, desto bättre, mer initierade och effektiva beslut kan organisationen ta. I avsnitt 2.1.2 nämns fyra olika metoder för underhållsplanering för ett ledningsnät. En ANN-modell behöver inte ersätta dessa, utan kan fungera som ett komplement.

Resultatet från ANN-modellen är tvåfaldigt: modellen ger dels en indikation på läcka, dels erhålls prediktioner mellan noll och ett för varje ledningsobjekt. Den första delen kan används för planerat underhåll:

ledningar med risk för läckage identifieras och kan åtgärdas innan ett akutärende uppstår.

Den senare delen, prediktionerna, knyter snarare an till kondition och prestation, och kan används för proaktivt underhåll. Prediktionerna är inte ett mått på kondition, men ju närmare ett en prediktion är, desto säkrare är modellen på ledningen bör klassas som ett. Eftersom läcka kan vara ett resultat av dålig kondition kan därför en hög prediktion tolkas som att vissa förutsättningar för läcka är uppnådda, och därför finns det risk för att konditionen är dålig. Prediktionerna möjliggör därför identifiering av områden där inga läckor nödvändigtvis uppstått, men där modellen visar en förhöjd risk för läckage. Detta möjliggör en mer långsiktig planering där områden med bristande kondition kan identifieras och prioriteras vid underhållsplanering.



Figur 36. De förhållningssätt till underhåll där en ANN-modell kan tillämpas är inringade. Omarbetat från Jacobsson m. fl. (2019).

En ANN-modell utvärderar ledningsnätet på objektnivå, men det är viktigt att komma ihåg att trots den skenbara noggrannheten med utvärdering både på enskilda objekt och en kontinuerlig klassning av risken för läcka från noll till ett, är det bara en modell och inte ett facit.

Innan en mer utförlig utvärdering gjorts angående hur väl ANN-modeller presterar på enskilda objekt i praktiken, bör resultatet från en ANN-modell endast användas som en indikationen över ledningsnätets kvalitet och var ytterligare utredningar bör göras – inte som ett mått på vilka ledningar som ska bytas ut.

Sverige består av 290 kommuner och i många av dessa är resurserna små och det är svårt att arbeta planerat med ledningsnätet. En viktig aspekt av ANN-modellen är att den är relativt enkel att köra när indata sammanställts. Därför kan den appliceras i kommuner där resurserna för underhåll är små. Eftersom denna studie indikerat att ANN-modellen kan köras med enkla och generella attribut kan en ANN-modell därför vara ett viktigt beslutsstöd för dessa kommuner och möjliggöra en bättre underhållsplanering.

5.6 Analys av prediktionerna

Eftersom orsaksfaktorer till brott varierar mellan olika material delades många analyser upp materialvis. I analysen ingick fyra material vilket innebar att för varje attribut som analyserades bildas $4 \times n_{kat}$ grupper, där n_{kat} är antalet unika kategorier för ett attribut, vilket medförde att ifall attributet som analyserades hade många kategorier blev det många grupper, och antalet prover i respektive grupp kunde därför vara få. Detta kan medföra att vissa kategoriers läckfrekvenser inte är särskilt säkra. För att få säkrare analyser vore det önskvärt med mer data.

Gamla ledningar har högre läckfrekvens än yngre (figur 22) och den kraftiga ökningen som kan ses efter en ålder på 45 år indikerar att någonting förändrades. Tabell 15 visar medelåldern för de fyra vanligaste materialen. Högst medelålder har gjutjärn med en ålder på 66 år, medan medelåldern för det näst äldsta materialet, segjärn, är 46 år. Att läckfrekvensen ökar när gjutjärn lades mest frekvent, tyder på att läckfrekvensen reducerats då andra material än gjutjärn började användas. Gjutjärn har dessutom mycket högre läckfrekvens än de andra materialen, illustrerat i figur 21. Korrosion, belastning och delvis sättning är faktorer där effekten ökar med tiden, vilket kan vara bidragande till den med åldern ökande läckfrekvensen.

Tabell 15. Medelålder för respektive material.

Material	Ålder
GJJ	66
SEG	46
PVC	43
PE	20

För gjutjärn är läckfrekvensen för gruppen med minst diameter markant lägre än för de två efterföljande (figur 23) vilket inte motsvarar det förväntade enligt Sundahl (1996). I indata fanns endast 11 gjutjärnsledningar med en diameter på 61 mm eller mindre. Detta kan vara en förklaring till denna diskrepans mellan verklighet och teori.

Analysen av trafiklast visade att för gjutjärn var cykelväg och öppen mark klart överrepresenterade i läckfrekvensen (figur 25). Detta är lite märkligt – intuitivt bör högre laster ge fler läckor, vilket observerats i tidigare studier för gjutjärn (Sægrov, 1998). En förklaring när det gäller öppen mark skulle kunna vara rötter (Sundahl, 1996), men analysen av markanvändning (figur 28) påvisade inte att skog skulle vara överrepresenterat i läckfrekvens. För att utvärdera om rotpåverkan är ett problem i Umeå behövs ett bättre mått än skog men något sådant attribut var inte tillgängligt för Umeå. För gjutjärn är det endast 88 ledningar som ligger i antingen öppen mark eller cykelväg (58 respektive 30 stycken), vilket kan innebära att läckfrekvenserna egentligen inte är representativa. För de andra materialen uppvisas oftast vad som förväntas: högra laster ger högre läckfrekvens.

Berg är med marginal överrepresenterat för både gjutjärn och PVC när jordmaterial undersökts. Det är dock endast ett fåtal ledningar med klassen Berg och därför bör ingen vikt läggas vid detta. Vidare förekommer endast sammanlagt 120 ledningar med klassen torv så trots att torv har lägst läckfrekvens för alla material förutom PVC innebär det inte att det är fördelaktigt att lägga ledningar i torv. Friktionsjord har hög läckfrekvens för gjutjärn. Eftersom friktionsjord håller kvar vatten sämre än kohesionsjord är det enklare att packa friktionsjord än kohesionsjord vilket bör minska risken för sättningar. Dessutom är kohesionsjord mer

tjälfarlig än friktionsjord. Detta talar för att friktionsjord borde vara ett bättre material än kohesionsjord att lägga gjutjärnsledningar i, men ANN-modellen visar det motsatta. Orsaker till detta kan vara flera; i analysen är det många jordarter med olika kornstorlek, bildningssätt och så vidare som grupperats som friktionsjord (tabell 6) och därför kan det vara stor skillnad mellan jordar klassade som friktionsjord. Dessutom baseras jordartsanalysen från en kartlager från SGU som de avråder från att använda till analys då klassningen är väldigt grov. Det kan också vara så att jordarter grupperas i kluster så att attributet Jordmaterial fångar in andra faktorer än just jordmaterial. Vidare finns det faktorer som inte beaktats i den här modellen, exempelvis packningsgrad, utformning av ledningsbädd och fyllningsmaterial. För segjärn är kohesionsjord överrepresenterat även om både friktionsjord och morän erhåller relativt höga värden – detta kan bero på förhöjd korrosionsrisk i kohesionsjord. Detsamma gäller ifall moränen innehåller mycket lera. För PVC är läckfrekvensen för friktionsjord och kohesionsjord mer än dubbelt så hög som läckfrekvensen för morän. Moräns egenskaper varierar med vilka jordarter som ingår i moränen och det är därför svårt att utvärdera morän utan en tydligare klassning. För PE är kohesionsjord överrepresenterat i läckfrekvensen. Dimensionsändring har främst identifierats som en viktig faktor för brott på PE (Malm, Horstmark, Larsson m. fl., 2011) och kohesionsjords överrepresentation är därför svår att förklara. PE är segare än de andra materialen, men trots detta skulle sättningar kunna vara en förklaring till denna överrepresentation i läckfrekvens.

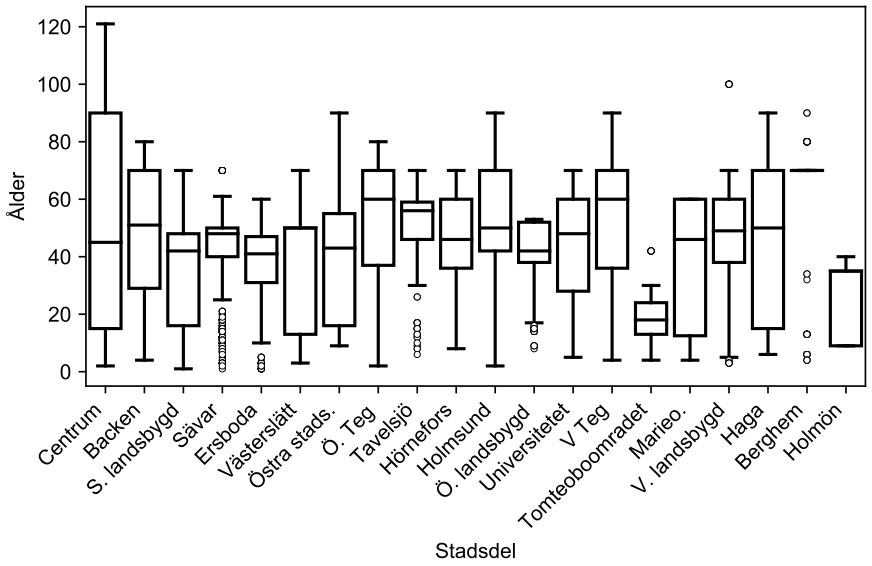
För tryckzon är lågzonen klart överrepresenterat för läckor (figur 27). Eftersom tryckzon är områdesbaserat blir det snarare ett indirekt attribut än ett direkt – attributet fångar upp andra faktorer än endast tryckzon. Tryckzon skulle kunna beskriva tryckskillnader i ledningsnätet, men i Umeå är det inte känt att det skulle vara ett högre medeltryck i någon av zonerna – det skulle dock kunna vara en förklaring. Att ha med tryck som attributet i ANN-modellen vore intressant men det var inte möjligt att inkludera i den här studien. En viktig del av attributet är att det markerar två skilda zoner så det fångar upp eventuella skillnader mellan dessa. Problemet är att det också fångar upp andra skillnader såsom jordarter, ålder, befolkning och så vidare. Den stora skillnaden

mellan de två tryckzonerna behöver därför inte bero på att det är skilda zoner, utan kan bero på andra faktorer. Tryckzon korrelerar inte starkt med något av de andra attributen (utöver Stadsdel som också är ett indirekt attribut) så det finns ingen tydlig förklaring till den stora skillnaden i läckfrekvens. Tryckzon visar ändå relativt hög korrelation mot Stadsdel (0,9), Jordmaterial (0,3) och Befolkningsförändring (0,3), se figur 17. Jordmaterial och Befolkningsförändring skulle kunna bidra till skillnaden i läckfrekvens, men samtidigt är korrelationen låg. Ytterligare en förklaring kan vara att Tryckzon korrelerar mot exempelvis flera dåliga material. Korrelationen mot enskilda material blir då låg, men korrelationen skulle kunna vara hög mot en grupp av material. Det kan också vara flera attribut som samverkar men som inte syns i de korrelationsanalyser som gjorts. Detta behöver undersökas innan det går att konstatera att Umeås lågzon har en inneboende egenskap som ökar läckfrekvensen för tillhörande ledningar.

Markanvändning uppvisar vad som intuitivt kan förväntas: områden med höga laster har ökad läckfrekvens (figur 28). Intressant är att attributet Järnvägs påverkan på modellens noggrannhet har visat sig vara försumbart, men markanvändningen Järnväg uppvisar ökad läckfrekvens jämfört med skog och grönyta, men även urban miljö. En anledning till denna diskrepans kan vara den låga förekomsten av ledningssträckor med markanvändningen Järnväg: 33 stycken. Anledningen till att två olika mått för järnvägs påverkan finns är för att Järnväg i Markanvändning endast omfattar spårens utbredning. Eftersom vibrationer och laster har en lateral utbredning, skapades ett attribut där även ledningar i anknytning till järnväg inkluderades. Trots den bredare definitionen är det endast 65 ledningsobjekt som inkluderas i attributet Närliggande järnväg. För att få en bättre bild av järnvägs påverkan skulle fler objekt påverkade av järnväg behöva analyseras. I RörANN-projektet testas nu att köra en modell med data från olika delar av Sverige – i den analysen är inte järnväg inkluderad, men detta förfarande skulle kunna möjliggöra en bättre utvärdering av Järnväg.

Stadsdelarna uppvisar tydlig skillnad i läckfrekvens (figur 29). Stadsdel är ett indirekt attribut och det är underliggande faktorer som ligger bakom skillnaden i läckfrekvens mellan stadsdelarna. Analysen av ålder

(figur 22) visar tydligt att läckfrekvensen är högre för äldre ledningar. MLR-analysen påvisar dock ingen korrelation mellan stadsdel och ålder (figur 18). Lådagrammet i figur 37 åskådliggör att medianåldern skiljer sig mellan stadsdelar, men stora överlapp mellan de olika stadsdelarnas åldersfördelning förekommer. Dessutom är det tydligt i figur 29 att det inte är de stadsdelar med högst ålder som har högst läckfrekvens. Den kategoriska korrelationsanalysen visar att stadsdel korrelerar starkt med attributen Tryckzon och Befolkningsförändring (figur 17). Vidare har Stadsdel en korrelation på 0,3 eller högre med Jordmaterial (0,4), Trafiklast (0,3), Material (0,4), Markanvändning (0,3), Fjärrvärme (0,3) och Innerdiameter (0,3). Detta är låga korrelationsvärden, men de flesta kombinationer av attribut har korrelationsvärden under 0,3. Skillnaden i läckfrekvens mellan stadsdelar kan därför vara ett samlat utfall av dessa attribut.



Figur 37. Ett lådagram mellan attributen Stadsdel och Ålder. Inom respektive låda ingår 50 procent av alla ledningsobjekt för varje stadsdel. Sträcket inom varje låda visar respektive stadsdels medianålder.

Att korsande fjärrvärme har dubbelt så hög läckfrekvens är inte konstigt (figur 30). Antalet läckor per år varierar mycket beroende på

bland annat väder och i Umeå är tjäle ett problem för ledningar. Runt fjärrvärmeledningar är temperaturen högre och detta bör rimligen leda till att jordrörelsen runt fjärrvärmeledningen blir mindre, vilket blir ett problem vid övergången mellan tjälad och otjälad jord. Detta skapar en liknande effekt som vid sättningar.

I figur 31 visas det att läckfrekvensen är högre för ledningar utan serviskopplingar. Rimligtvis bör varje koppling på en ledning introducera en svag punkt på ledningen och därför är det uppseendeväckande att läckfrekvensen är högre för ledningar utan serviskopplingar. En förklaring till att läckfrevensen på ledningar med serviskopplingar är lägre kan vara hur läckan rapporteras. I den här rapporten har läckor på distributionsledningar sökts ut och inte på servisledningar. Det är möjligt att många läckor vid serviskopplingar hänförts till servisledningen och inte huvudledningen. Läckor har även kopplats till vattenledningar genom att buffra kring ledningen och läckor nära huvud- och distributionsledningar borde därför ha fångats upp, men ibland är inte läckorna markerade precis där de skedde och därför kan de ha missats.

Denna utvärdering har visat att ANN-modellen många gånger motsvarar tidigare forskning vilket indikerar att modellen identifierat viktiga samband. Samtidigt finns det läckfrekvenser som sticker ut och som inte är vad som förväntas. Detta kan innebära att modellen missat vissa samband, eller så förekommer andra förklaringsmodeller som inte identifierats i det här arbetet.

5.7 Vidare arbete

Resultatet från det här arbetet indikerar att antalet attribut i en ANN-modell kan begränsas betydligt utan att påverka noggrannheten i alltför hög grad. Vidare visar resultaten att det kan vara möjligt att köra en ANN-modell på kommuner med dåligt dataunderlag, men där blir det en kompromiss mellan noggrannhet för modellen och de attribut som kan uppbringas. Modellen bör köras på Vindelns kommun för att utvärdera möjligheten att köra en ANN-modell på en mindre kommun i praktiken.

En intressant utvärdering av modellens prestation skulle kunna erhållas ifall modellen tränas på hur ledningsnätet var exempelvis 2010

– det vill säga att modellen körs som att det var 2010 och inte som att det var 2020 (året när den här studien utfördes). Det skulle då gå att jämföra ifall de ledningar som modellen klassade som hög sannolikhet för läckage 2010 har lett till läckage 2020, och därefter utvärdera modellen utifrån detta. Nackdelen med detta är att det minskar datamängden, och mängden data var redan begränsad i den här studien. Detta skulle dock vara intressant för en kommun med mer data än Umeå kommun.

Som nämnts tidigare i diskussionen vore det också intressant att kunna köra en modell där facit är korrekt, men träna modellen på ett modifierat facit så att facit innehåller fel. Därefter kan prediktionen från modellen tränad på ett modifierat facit jämföras med en modell tränad på korrekt facit, alternativt direkt mot facit. Detta är så klart svårt, eftersom kvaliteten på ett dricksvattennät är svårt att undersöka – det är därför konditionen önskas uppskattas med en ANN-modell, men det kan likväl vara något att sträva mot i framtiden. Det kan arbetas mot genom att exempelvis notera kvaliteten på ledningar som byts ut i samband med vägarbete och dylikt.

Något som är intressant är att det i Umeå kommun placerats ut vattenmätare som lyssnar efter vattenläckor och som i realtid indikerar var det finns risk för vattenläckor. Dessa läckor är begränsade till att ligga nära vattenmätarna, men det är ändå intressant att kunna jämföra resultat från modellen mot en ögonblicksbild av vattennätet. Genom att exportera indikerade läckor från vattenmätarna och importera dem till ett GIS-program och överlagra dem på ledningsnätet skulle det gå att jämföra var ANN-modellen och vattenmätarna är överens. Detta är inte svårt, men just nu finns endast dessa vattenmätare i ett närliggande samhälle och inte i Umeå stad, och därför skulle en sådan jämförelse inte bli särskilt omfattande. Information från mätarna skulle också kunna användas i träningsskedet för att exempelvis klassa en ledning som i behov av underhåll ifall en mätare påvisar läcka. Även detta skulle dock kräva fler mätare, och mätare i Umeå, för att det skulle vara givande. Det är i alla fall något som är intressant för framtiden, och något som kan bli aktuellt i andra kommuner också.

6 Slutsatser

Det råder en viss osäkerhet kring modellens prestation. En noggrannhet på 0,8 kan inte endast hänföras till osäkerhet i facit. Modellen presterar bättre än slumpen, men skulle behöva förbättras för att få mer trovärdiga resultat. Detta kan uppnås med mer data, eller en mer komplicerad modell. Ju högre tröskel som väljs för läcka, desto större andel historiska läckledningar identifierar modellen. Detta kan tolkas som att modellen blir säkrare och säkrare ju högre tröskelvärde och denna egenskap kan utnyttjas då ledningsnätet utvärderas.

I den här studien gjordes avgränsningen att endast fokusera på huvudledningar och distributionsledningar och inte servisledningar. Detta för att bland annat olika flödessituationer förekommer mellan dessa ledningar och servisledningar. Att inkludera servisledningar skulle kunna öka datamängden och det skulle därför kunna vara en bra idé att inkludera servisledningar i modellen också. För att markera skillnaderna mellan ledningarna skulle en dummy-variabel kunna användas för att klassa ledningar som ej servisledningar eller servisledningar.

Även med en stor reduktion av antalet attribut kunde en noggrannhet uppnås som nästan motsvarade noggrannheten då alla attribut inkluderas. Med tio attribut uppnås en noggrannhet på 0,79, att jämföra med en noggrannhet på 0,8 då alla attribut är inkluderade. De tio attributen var Jordart, Tryckzon, Markanvändning, Befolkningsförändring, Stadsdel, Trafiklast, Innerdiameter, Material, Fjärrvärme och Serviskoppling (tabell 10).

Med attribut som kan vara enkla att samla in och som generaliserar väl (figur 8) uppnåddes en noggrannhet på 0,75. Detta skiljer sig inte mycket från hur modellen presterar med alla attribut. Detta indikerar att modellen även kan anpassas på kommuner med sämre tillgång till data. Att applicera modellen på Vindelns kommun kan bli svårt då osäkerhet råder kring viktiga attribut såsom Ålder, Innerdiameter och Material, men detta behöver undersökas ytterligare.

Viktiga attribut analyserades med ReliefF, RFC-RFECV, MLR-RFECV och korrelation. Både ReliefF och MLR-RFECV presterade väl. Med de tio viktigaste attributen enligt ReliefF presterade modellen

nästan lika väl som modellen med alla attribut. ANN-modellen presterade jämförbart mellan de sju viktigaste attributen enligt ReliefF respektive MLR-RFECV. RFC-RFECV lyckades endast urskilja tre attribut som mindre viktiga och därför användes inte RFC-RFECV för attributurval. Det finns bättre sätt att applicera RFC-RFECV än vad som gjorts i den här studien och ett annat resultat hade då kunnat erhållas. Inte heller korrelation kunde användas i någon större utsträckning för attributurval.

Resultatet från en ANN-modell kan användas för strategiskt underhåll. Eftersom prediktionerna ges objektsvis är det möjligt att identifiera enskilda ledningsobjekt med hög risk för läcka innan en stor läcka uppstår. Att prediktionerna är kontinuerliga kan dessutom användas för att dela in ledningsnätet i olika riskklasser. En spatial analys kan visa att många ledningar inom ett område har en prediktion inom ett visst intervall, och lämpliga åtgärder kan då planeras utifrån det. Prediktionerna är inget facit, men de kan indikera var ytterligare utredningar ska göras. Andra metoder som i dag används som beslutsstöd för underhåll är ofta översiktliga och baserade på tumregler, och en ANN-modell kan då fungerar som ett komplement

Att modellen presterar jämförbart mellan en fullständig modell och en modell tränad på data tillgängliga för kommuner med sämre tillgång till data påvisar att en ANN-modell kan användas för strategiskt underhåll även i sådana kommuner. Det är relativt enkelt att applicera en ANN-modell, och om modellen kan tränas på en närliggande kommun med mycket data kan modellen köras även då mängden historiska data är liten.

Resultatet från ANN-modellen har jämförts med tidigare forskning med avseende på faktorer som påverkar läckfrekvens. I många aspekter motsvarar resultatet vad som kan förväntas. Exempelvis för PE är är läckfrekvensen markant högre där dimensionsändring sker (figur 24) och läckfrekvensen ökar generellt med minskande diameter (figur 23). I andra fall är resultatet inte vad som kan förväntas. Läckfrekvensen för gjutjärn (figur 26) i friktionsjord är exempelvis mycket högre än för någon annan sorts jordmaterial och detta är något som måste undersökas noggrannare.

Referenser

Acock, A. C. & Stavig, G. R. (1979). A measure of association for nonparametric statistics. *Social Forces*.

Aldehim, G. & Wang, W. (2015). Determining appropriate approaches for using data in feature selection. *International Journal of Machine Learning and Cybernetics*, *8*(3), 915–928. https://doi.org/10.1007/s13042-015-0469-8

Beretta, L. & Santaniello, A. (2011). Implementing ReliefF filters to extract meaningful features from genetic lifetime datasets. *Journal of Biomedical Informatics*, *44*(2), 361–369. https://doi.org/10.1016/j.jbi.2010.12.003

Bergsma, W. (2012). A bias-correction for cramérs v and tschiprow's t. *Journal of the Korean Statistical Society*.

Brungard, C. W., Boettinger, J. L., Duniway, M. C., Wills, S. A. & Edwards, T. C. (2015). Machine learning for predicting soil classes in three semi-arid landscapes. *Geoderma*, *239-240*, 68–83. https://doi.org/10.1016/j.geoderma.2014.09.019

Castillo, P., Merelo, J., Prieto, A., Rivas, V. & Romero, G. (2000). G-prop: Global optimization of multilayer perceptrons using GAs. *Neurocomputing*, *35*(1-4), 149–163. https://doi.org/10.1016/s0925-2312(00)00302-7

Chollet, F. m. fl. (2015). Keras.

Chollet, F. (2017, 28. oktober). *Deep learning with python*. Manning Publications. https://www.ebook.de/de/product/28930398/francois_chollet_deep_learning_with_python.html

Clark, R. M., Allen, M. & O'Day, D. K. (1987 juni). *Water main evaluation for rehabilitation/replacement* (tekn. rapport). United States Environmental Protection Agency.

Finansdepartementet. (2020, 23. april). *Statens budget i siffror* (Regeringskansliet, Red.). https://www.regeringen.se/sveriges-regering/finansdepartementet/statens-budget/statens-budget-i-siffror/

Geografiska Sverigedata. (2016, 1. januari). Gsd-terrängkartan, raster (Lantmäteriet, Red.).

Gregorutti, B., Michel, B. & Saint-Pierre, P. (2016). Correlation and variable importance in random forests. *Statistics and Computing*, *27*(3), 659–678. https://doi.org/10.1007/s11222-016-9646-1

Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*.

Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1/3), 389–422. https://doi.org/10.1023/a:1012487302797

Hakkarainen, M. & Albertsson, A.-C. (2004). Environmental degradation of polyethylene. *Long term properties of polyolefins* (s. 177–200). Springer Berlin Heidelberg. https://doi.org/10.1007/b13523

Hall, M. A. (1999 april). *Correlation-based feature selection for machine learning* (doktorsavhandling). The University of Waikato.

Jacobsson, D., Giertz, T. & Adrup, A. (2019, 26. september). *Effektivt underhåll av va-system (remissversion)* (tekn. rapport). Svenskt Vatten.

Jafar, R., Shahrour, I. & Juran, I. (2010). Application of artificial neural networks (ANN) to model the failure of urban water mains. *Mathematical and Computer Modelling*, *51*(9-10), 1170–1180. https://doi.org/10.1016/j.mcm.2009.12.033

Keras. (2019, 16. december). *Dense* (K. Documentation, Red.). https://keras.io/layers/core/#dense

Kim, G., Kim, Y., Lim, H. & Kim, H. (2010). An MLP-based feature subset selection for HIV-1 protease cleavage site analysis. *Artificial Intelligence in Medicine*, *48*(2-3), 83–89. https://doi.org/10.1016/j.artmed.2009.07.010

Kira, K. & Randell, L. A. The feature selection problem: Traditional methods and a new algorithm. I: *Aaai-92 proceedings*. 1992.

Kononenko, I., Robnik-Sikonja, M. & Pompe, U. (2000 februari). Relieff for estimation and discretization of attributes in classification, regression, and ilp problems.

Kononenko, I., Šimec, E. & Robnik-Šikonja, M. (1997). Overcoming the myopia of inductive learning algorithms with relieff. *Applied*

Intelligence, *7*(1), 39–55. https://doi.org/10.1023/a:1008280620621

Kutyłowska, M. (2016). Comparison of two types of artificial neural networks for predicting failure frequency of water conduits. *Periodica Polytechnica Civil Engineering*. https://doi.org/10.3311/ppci.8737

Lidström, V. (2013 augusti). *Vårt vatten: Grundläggande lärobok i vatten- och avloppsteknik*. Svenskt Vatten.

Malm, A., Horstmark, A., Jansson, E., Larsson, G., Meyer, A. & Uuijärvi, J. (2011). *Handbok i förnyelseplanering av va-ledningar* (tekn. rapport Nr 2011-12). Svenskt Vatten.

Malm, A., Horstmark, A., Larsson, G., Uusijärvi, J., Meyer, A. & Jansson, E. (2011). *Rörmaterial i svenska va-ledningar—egenskaper och livslängd* (tekn. rapport Nr 2011-14). Svenskt Vatten.

Malm, A. & Svensson, G. (2011). *Material och åldersfördelning för sveriges va-nät och framtida förnyelsebehov* (tekn. rapport Nr 2011-13). Svenskt Vatten.

MathWorks. (2020, 28. april). *Jbtest*. https://www.mathworks.com/help/stats/jbtest.html

McDonald, J. H. (2014). *Handbook of biological statistics* (3. utg.). Sparky House Publishing.

McKinney, W. Data Structures for Statistical Computing in Python (S. van der Walt & J. Millman, Red.). I: I *Proceedings of the 9th Python in Science Conference* (S. van der Walt & J. Millman, Red.). Utg. av van der Walt, S. & Millman, J. 2010, 56–61. https://doi.org/10.25080/Majora-92bf1922-00a.

Nisbet, R., Yale, K. & Minere, G. (2018). *Handbook of statistical analysis and data mining applications*. Elsevier. https://doi.org/10.1016/c2012-0-06451-4

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Rajani, B. & Kleiner, Y. (2001). Comprehensive review of structural deterioration of water mains: Physically based models. *Urban Water*, *3*(3), 151–164. https://doi.org/10.1016/s1462-0758(01)00032-2

Rehn, D. & Giertz, T. (2019, 11. januari). *En ai-modell för vattenledningsnätet* (tekn. rapport). Stockholm Vatten och Avfall.

Robnik-Šikonja, M. & Kononenko, I. (2003). Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning*, *53*(1), 23–69. https://doi.org/10.1023/A:1025667309714

Rodgers, J. L. & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*.

Sægrov, S. (1998, 3. mars). *Forfall oc fornyelse av ledningsnett* (tekn. rapport). Norsk VA-verkforening.

Salesi, S., Alani, A. A. & Cosma, G. A hybrid model for classification of biomedical data using feature filtering and a convolutional neural network. I: *2018 fifth international conference on social networks analysis, management and security*. 2018.

Seabold, S. & Perktold, J. Statsmodels: Econometric and statistical modeling with python. I: *9th python in science conference*. 2010.

Shi, S., Li, G., Chen, H., Liu, J., Hu, Y., Xing, L. & Hu, W. (2017). Refrigerant charge fault diagnosis in the VRF system using bayesian artificial neural network combined with ReliefF filter. *Applied Thermal Engineering*, *112*, 698–706. https://doi.org/10.1016/j.applthermaleng.2016.10.043

SPSS. (2020, 8. april). *Chi-square goodness-of-fit test – simple tutorial*. https://www.spss-tutorials.com/chi-square-goodness-of-fit-test/

Srinivas, N. S. S., Sugan, N., Kar, N., Kumar, L. S., Nath, M. K. & Kanhe, A. (2019). Recognition of spoken languages from acoustic speech signals using fourier parameters. *Circuits, Systems, and Signal Processing*, *38*(11), 5018–5067. https://doi.org/10.1007/s00034-019-01100-6

Stahre, P., Mellström, G. & Adamsson, J. (2007). *Värdering av vatten- och avloppsledningsnät* (tekn. rapport Nr 2007-13). Svenskt Vatten.

Stål, T. & Wedel, P. (1984). *Handboken bygg. g, geoteknik*. Liber.

Studenmund, A. H. (2011). *Using econometrics: A practical guide* (6. utg.). Addison-Wesley.

Sun, Y. (2007). Iterative RELIEF for feature weighting: Algorithms, theories, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(6), 1035–1051. https://doi.org/10.1109/tpami.2007.1093

Sundahl, A.-C. (1996). *Diagnos av vattenledningars kondition* (Licentiatuppsats). Institutionen för Teknisk Vattenresurslära. Lunds Tekniska Högskola. Lunds Universitet.

The pandas development team. (2020). *Pandas-dev/pandas: Pandas* (Version latest). Zenodo. https://doi.org/10.5281/zenodo.3509134

Thi, H. A. L. & Nguyen, M. C. (2016). DCA based algorithms for feature selection in multi-class support vector machine. *Annals of Operations Research*, *249*(1-2), 273–300. https://doi.org/10.1007/s10479-016-2333-y

Urbanowicz, R. J., Meeker, M., Cava, W. L., Olson, R. S. & Moore, J. H. (2018). Relief-based feature selection: Introduction and review. *Journal of Biomedical Informatics*, *85*, 189–203. https://doi.org/10.1016/j.jbi.2018.07.014

Urbanowicz, R. J., Olson, R. S., Schmitt, P., Meeker, M. & Moore, J. H. (2018). Benchmarking relief-based feature selection methods for bioinformatics data mining. *Journal of Biomedical Informatics*, *85*, 168–188. https://doi.org/10.1016/j.jbi.2018.07.015

VASS. (2020, 23. april). *Definition av vattenläcka* (Svenskt Vatten, Red.). http://www.vass-statistik.se/

Walan, P. (2019, 9. oktober). *Underlag till förnyelseplan för vattenledningsnätet 2019 – 2023* (tekn. rapport). Vakin.

Appendix A: Datautvinning

I tabell A1 redovisas kort hur data för respektive attribut samlats in. MyCarta Analyze, som använts för att extrahera viss data, fungerar inte som ett typiskt GIS-program och beskrivs därför kort i följande stycke.

I MyCarta Analyze väljs först ett startlager, exempelvis ett vattenledningsnät, och alla efterföljande operationer görs sedan på detta startlager. Om exempelvis startlagret buffras för att kunna inkludera punkter, kommer informationen från punktera att kopplas till startlagret, men efter operationen återgår startlagret till sin ursprungliga form; i fallet vattenledning återgår alltså ledningen från en buffrad polygon till en linje, men med ytterligare data kopplat till sig. Detta skiljer sig alltså från typiska GIS-program som ArcGIS och QGIS, där den buffrade vattenledningen skulle ha sparats till ett nytt kartlager i form av en polygon.

Tabell A1. Hur data samlats in för respektive attribut.

Attribut	Hur data tagits fram
Anläggnings-år	Fanns i Vakins databas.
Ålder	Beräknades enligt: Nutid – Anläggningsår.
Innerdiameter	För de flesta material, exklusive PE, fanns innerdiameter angivet i Vakins databas. För PE fanns endast ytterdiameter. För att konvertera detta till innerdiameter användes Pipelifes produktkatalog.[1]
Antal serviser	I MyCarta Anlyze, ett analysprogram till VA-banken som möjliggör spatiala analyser, gjordes en rumslig analys där det identifierades ifall en servisledning skar en ledning eller ej. Om så var fallet tilldelades ledningen en etta, annars en nolla.

Fortsätter på nästa sida

Tabell A1 – *Fortsättning*

Attribut	Hur data tagits fram
Antal ventiler	I MyCarta Analyze användes vattenledningsnätet som startlager och servisledningar filtrerades bort. Därefter tilldelades varje ledningsobjekt en tvåa ifall ledningsobjektet hade avstängningsventiler i båda ändar, en etta ifall endast i en ände och en nolla ifall ledningen saknade avstängningsventiler. För att förenkla för modellen klassades ledningar med minst en ventil till ett, annars noll.
Material	Ledningsmaterial identifierades för respektive ledningssträcka. De ledningar med ovanliga material sorterades bort och kvar blev gjutjärn, segjärn, PVC och PE.
Jordlager	För Umeå fanns det tillgängligt ett kartlager med jordarter. De olika jordarterna grupperades för att minska antalet kategorier (se tabell 6). Därefter överlagrades dessa grupper på ledningsnätet.
Berggrund	För Umeå kommun fanns det tillgängligt ett kartlager med berggrund. Dels angavs bergart, dels angavs bergartsgrupp. Att ange bergart är mer detaljerat, men bergartsgrupp valdes för att göra modellen mindre komplicerad. De fyra förekommande bergartsgrupperna var basisk plutonit/metaplutonit, sedimentär bergart, basisk vulkanit och sur-intermediär plutonit/metaplutonit.
Markanvändning	För Umeå kommun fanns det tillgängligt ett kartlager där markanvändning redovisades. För att minska komplexiteten reducerades antalet kategorier enligt tabell 7.

Fortsätter på nästa sida

Tabell A1 – *Fortsättning*

Attribut	Hur data tagits fram
Trafiklast	Umeå kommun hade kartlager med europavägar, större vägar, bilvägar, cykelvägar och gångvägar. Vägarna var angivna som linjer och för att representera vägarnas bredd buffrades de till olika storlekar enligt: europavägar 7,5 meter, större vägar 6,5 meter, busslinje 5 meter, bilvägar 5 meter, cykelväg 2,5 meter och gångväg 2 meter. De olika vägarna sammanfogades därefter och överlagrades på ledningsnätet. Vissa av vägarna överlappade varandra, och ibland gick en ledningssträcka från en vägtyp till en annan. I sådana fall, där en ledningssträcka blivit tilldelad olika vägtyper, behölls endast information om den största vägen.
Befolknings-förändring	Umeå kommun har sammanställt befolkningsförändring för olika stadsdelar under perioden 2005 – 2018. De stadsdelar där befolkningen ökat med mer än 20 procent tilldelades 1, de som minskat med mer än 20 procent tilldelades -1, och resterande områden tilldelades 0. Inga stadsdelar hade dock minskat med mer än 20 procent och därför förekom endast klasserna 0 och 1. Därefter överlagrades informationen på vattenledningsnätet.

Fortsätter på nästa sida

Tabell A1 – *Fortsättning*

Attribut	Hur data tagits fram
Dimensions-ändring	I MyCarta Analyze valdes vattenledningsnätets punkter som startlager, där punkter innefattar kopplingar, ventiler och så vidare. Till respektive punkt kopplades därefter alla ledningar vars tillpunkt (en ledningssträcka går mellan en tillpunkt och en frånpunkt i VA-banken) hade samma ID som startpunkten. Samma koppling gjordes sedan igen, men där ledningssträckors frånpunkt kopplades till startpunkten. När de ledningar som anslöt till startpunkten identifierats, jämfördes dimensionen på tillpunktsledningarna och frånpunktsledningarna i respektive punkt. Om de inte överensstämde tilldelades attributet en etta – det vill säga dimensionsändring förekommer.
Järnväg	I MyCarta Analyze laddades vattenledningsnätet som startdata. Därefter skapades en buffert på tio meter varpå järnvägsnätet överlagrades. Alla ledningar med järnväg inom tio meter tilldelades en etta.
Tryckzon	I Umeå finns två olika tryckzoner, och i Vakins databas angavs för de flesta vattenledningar vilken tryckzon ledningen tillhörde.
Höjd	Umeå kommun har ett raster där höjd anges för 2×2-meters rutor. Dessa överlagrades på ledningsnätet med funktionen *v.rast.stats* från GRASS-modulen i QGIS. Både min- och maxvärdet extraherades från rasterlagret.

[1] https://www.pipelife.se/se/

Appendix B: Korrelation

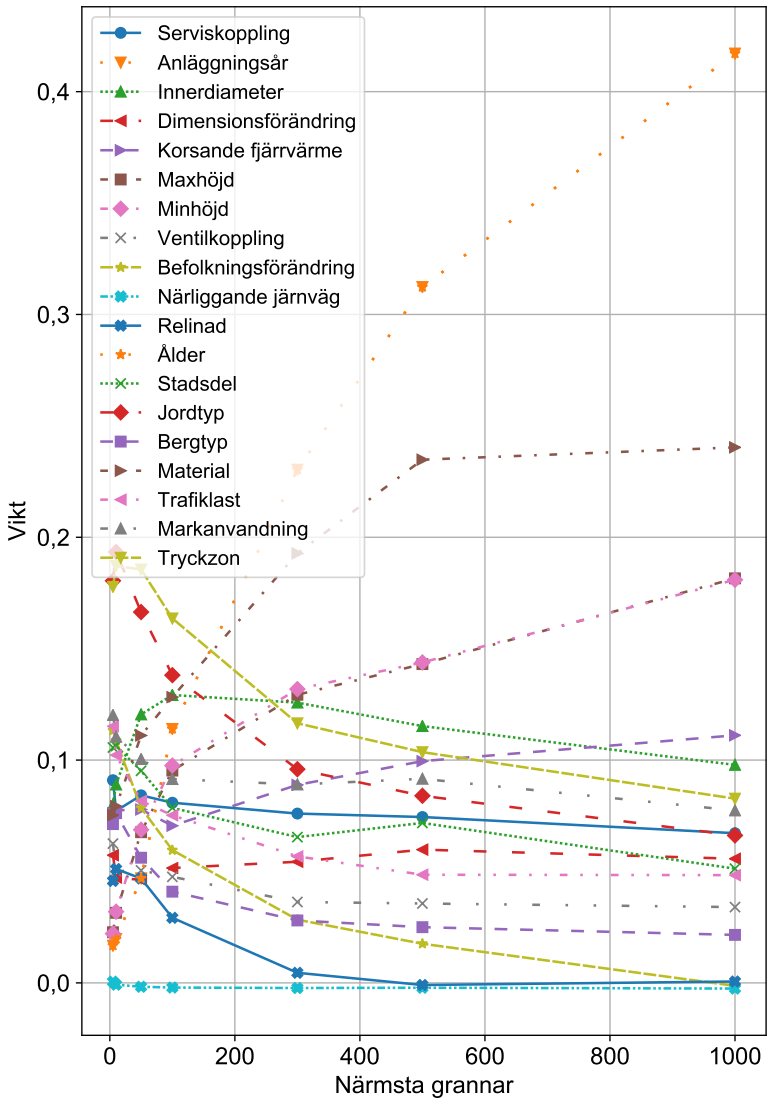
Tabell B1 visar Jarque Bera-värdet för residualerna från de olika kombinationerna av kvotdata. Vid höga JB-värden förkastas nollhypotesen att indata är normalförldelade. JB-värdet är väldigt högt för alla kombinationer av residualer och data kan därför inte antas vara normalfördelad.

Figur B1. Jarque Bera-värde för residualerna från de olika kombinationerna av kvotdata.

Variabler		
Oberoende	Beroende	Jarque-Bera-värde
Anl.	Diam.	129
	Maxhöjd	155
	Minhöjd	154
Diam.	Anl.	14 872
	Maxhöjd	15 574
	Minhöjd	15 559
Maxhöjd	Anl.	28 063
	Diam.	25 437
	Minhöjd	1 841 639
Minhöjd	Anl.	28 640
	Diam.	25 855
	Maxhöjd	1 644 654

Appendix C: ReliefF

När k går mot ett stort värde blir ReliefF-analysen snarare univariat eftersom närhetsaspekten försvunnit. Vid k=1000 fås ett helt annat resultat än vid k=5 eller k=10. Ålder och anläggningsår får då högst vikt, men material och höjd är också viktigt. Detta kan ses i figur C1.



Figur C1. Attributens vikter från ReliefF-algoritmen.

Tabell C1 visar modellens noggrannhet när den körs med fler och fler attribut, tillagda i den ordning som resultat från ReliefF-algoritmen indikerade. De första modellerna är dels körda med 68 noder vilket är det antal som den slutgiltiga modellen använde, samt ett antal noder som beror på antalet attribut som inkorporeras.

Tabell C1. ANN-modellen körd med resultatet från ReliefF-algoritmen där ett mindre viktigt attribut läggs till i varje ny modell. För att se effekten av antalet noder körs de första modellerna två gånger – en gång med få noder, en gång med många.

Mod.	Nod.	Min.	Med.	Max.	Attribut
1	68	0,50	0,56	0,60	Jordtyp
1	15	0,53	0,56	0,60	Jordtyp
2	17	0,52	0,57	0,60	Jordtyp, Tryckzon
2	68	0,53	0,56	0,60	Jordtyp, Tryckzon
3	23	0,56	0,59	0,62	Jordtyp, Tryckzon, Markanvändning
3	68	0,54	0,58	0,63	Jordtyp, Tryckzon, Markanvändning
4	24	0,57	0,60	0,64	Jordtyp, Tryckzon, Markanvändning, Befolkningsförändring
4	68	0,59	0,61	0,65	Jordtyp, Tryckzon, Markanvändning, Befolkningsförändring
5	44	0,61	0,63	0,68	Jordtyp, Tryckzon, Markanvändning, Befolkningsförändring, Stadsdel
5	68	0,58	0,62	0,67	Jordtyp, Tryckzon, Markanvändning, Befolkningsförändring, Stadsdel
6	47	0,65	0,68	0,72	Jordtyp, Tryckzon, Markanvändning, Befolkningsförändring, Stadsdel, Trafiklast

Fortsätter på nästa sida

Tabell C1 – *Fortsättning*

Mod	Nod.	Min.	Med.	Max.	Attribut
6	68	0,62	0,67	0,75	Jordtyp, Tryckzon, Markanvändning, Befolkningsförändring, Stadsdel, Trafiklast
7	51	0,67	0,73	0,79	Jordtyp, Tryckzon, Markanvändning, Befolkningsförändring, Stadsdel, Trafiklast, Innerdiameter
7	68	0,68	0,72	0,75	Jordtyp, Tryckzon, Markanvändning, Befolkningsförändring, Stadsdel, Trafiklast, Innerdiameter
8	68	0,72	0,75	0,79	Jordtyp, Tryckzon, Markanvändning, Befolkningsförändring, Stadsdel, Trafiklast, Innerdiameter, Material
8	55	0,73	0,76	0,78	Jordtyp, Tryckzon, Markanvändning, Befolkningsförändring, Stadsdel, Trafiklast, Innerdiameter, Material
9	68	0,72	0,77	0,83	Jordtyp, Tryckzon, Markanvändning, Befolkningsförändring, Stadsdel, Trafiklast, Innerdiameter, Material, Korsande fjärrvärme

Fortsätter på nästa sida

Tabell C1 – *Fortsättning*

Mod	Nod.	Min.	Med.	Max.	Attribut
9	56	0,74	0,78	0,82	Jordtyp, Tryckzon, Markanvändning, Befolkningsförändring, Stadsdel, Trafiklast, Innerdiameter, Material, Korsande fjärrvärme
10	68	0,74	0,78	0,80	Jordtyp, Tryckzon, Markanvändning, Befolkningsförändring, Stadsdel, Trafiklast, Innerdiameter, Material, Korsande fjärrvärme, Serviskoppling
10	57	0,76	0,79	0,83	Jordtyp, Tryckzon, Markanvändning, Befolkningsförändring, Stadsdel, Trafiklast, Innerdiameter, Material, Korsande fjärrvärme, Serviskoppling
11	68	0,73	0,77	0,79	Jordtyp, Tryckzon, Markanvändning, Befolkningsförändring, Stadsdel, Trafiklast, Innerdiameter, Material, Korsande fjärrvärme, Serviskoppling, Bergtyp

Fortsätter på nästa sida

Tabell C1 – *Fortsättning*

Mod	Nod.	Min.	Med.	Max.	Attribut
12	68	0,71	0,77	0,82	Jordtyp, Tryckzon, Markanvändning, Befolkningsförändring, Stadsdel, Trafiklast, Innerdiameter, Material, Korsande fjärrvärme, Serviskoppling, Bergtyp, Relinad
13	68	0,72	0,77	0,82	Jordtyp, Tryckzon, Markanvändning, Befolkningsförändring, Stadsdel, Trafiklast, Innerdiameter, Material, Korsande fjärrvärme, Serviskoppling, Bergtyp, Relinad, Ventilkoppling
14	68	0,78	0,81	0,85	Jordtyp, Tryckzon, Markanvändning, Befolkningsförändring, Stadsdel, Trafiklast, Innerdiameter, Material, Korsande fjärrvärme, Serviskoppling, Bergtyp, Relinad, Ventilkoppling, Dimensionsändring

Fortsätter på nästa sida

Tabell C1 – *Fortsättning*

Mod	Nod.	Min.	Med.	Max.	Attribut
15	68	0,72	0,79	0,81	Jordtyp, Tryckzon, Markanvändning, Befolkningsförändring, Stadsdel, Trafiklast, Innerdiameter, Material, Korsande fjärrvärme, Serviskoppling, Bergtyp, Relinad, Ventilkoppling, Dimensionsändring, Minhöjd
16	68	0,77	0,79	0,82	Jordtyp, Tryckzon, Markanvändning, Befolkningsförändring, Stadsdel, Trafiklast, Innerdiameter, Material, Korsande fjärrvärme, Serviskoppling, Bergtyp, Relinad, Ventilkoppling, Dimensionsändring, Minhöjd, Maxhöjd
17	68	0,76	0,81	0,85	Jordtyp, Tryckzon, Markanvändning, Befolkningsförändring, Stadsdel, Trafiklast, Innerdiameter, Material, Korsande fjärrvärme, Serviskoppling, Bergtyp, Relinad, Ventilkoppling, Dimensionsändring, Minhöjd, Maxhöjd, Anläggningsår

Fortsätter på nästa sida

Tabell C1 – *Fortsättning*

Mod	Nod.	Min.	Med.	Max.	Attribut
18	68	0,79	0,82	0,88	Jordtyp, Tryckzon, Markanvändning, Befolkningsförändring, Stadsdel, Trafiklast, Innerdiameter, Material, Korsande fjärrvärme, Serviskoppling, Bergtyp, Relinad, Ventilkoppling, Dimensionsändring, Minhöjd, Maxhöjd, Anläggningsår, Ålder
19	68	0,75	0,80	0,84	Jordtyp, Tryckzon, Markanvändning, Befolkningsförändring, Stadsdel, Trafiklast, Innerdiameter, Material, Korsande fjärrvärme, Serviskoppling, Bergtyp, Relinad, Ventilkoppling, Dimensionsändring, Minhöjd, Maxhöjd, Anläggningsår, Ålder, Järnväg