Budapesti Műszaki és Gazdaságtudományi Egyetem
Villamosmérnöki és Informatikai Kar
Méréstechnika és Információs Rendszerek Tanszék

Hibrid kvalitatív-kvantitatív transzfer tanulás és alkalmazása épületek energetikai tulajdonságainak becslésére

TDK dolgozat

Készítette:

Lengyel Nándor

Konzulens:

Dr. Kocsis Imre

2022

Tartalomjegyzék

Kivonat

Az energiahatékonyság biztosítása rendkívül fontos kérdéssé vált, mind pénzügyi, mind klímavédelmi szempontból. Az energiafogyasztás egy jelentős részét az épületek üzemeltetése tesz ki, ezért elengedhetetlen, hogy optimalizáljuk az épületek energiafelhasználását. Az optimalizációs folyamat fő lépése, hogy meghatározzuk az épület energiahatékonyságát, és a hozzá tartozó elvárt fogyasztást, ami viszonyítási alapként szolgál a tényleges fogyasztás vizsgálatakor.

Az épületek energiahatékonysági kimutatásának egyik legelterjedtebb formája az energetikai tanúsítvány, amely magába foglalja többek között az épület energetikai jellemzőit, valamint az épület elvárt fogyasztását. Azonban számos épület nem rendelkezik tanúsítvánnyal, és a tanúsítvány előállítási folyamata komplex számításokat, és esetenként helyszíni bejárást igényel.

A kutatásom általános célja, hogy kiváltsam - az energetikai tanúsítás során szükséges - magas humán ráfordítást egy statisztikai alapú megközelítéssel, így az elvárt energiafogyasztás becslése általánosítható és skálázható lesz. A statisztikai modell az épületek kvalitatív energetikai jellemzőin (pl. fal, fűtésrendszer minősége) és a hozzátartozó elvárt fogyasztáson alapul. Az energetikai tanúsítványokból - megfelelő átalakítások révén - kinyerhetőek a kvalitatív jellemzők és az elvárt fogyasztás érték. Az elvárt érték meghatározása során a kvalitatív jellemzőket egy energetikus szemrevételezés útján meg tudja határozni, így nincs szükség a részletes felmérés és számítás elvégzésére.

A kutatás egyik fő kihívása, hogy míg egyes országok rendelkeznek nyílt és kiterjedt tanúsítvány adatbázisokkal, addig más országok nem publikálnak megfelelő mennyiségű vagy minőségű adatot. A kutatásom fő fókusza, hogy transzfer-tanulás segítségével a kevés, hiányos adattal rendelkező országokban is meg lehessen határozni az elvárt fogyasztást statisztikai úton. A tudás átvitele nem triviális probléma, hiszen nemcsak az épületek eloszlása változó országonként, hanem különböző módszereket alkalmaznak az elvárt értékek kiszámítására. A transzfer-tanulás során, ha a forrás-, és céldomén logikailag számottevően eltér egymástól, akkor előfordulhat negatív transzfer is. A transzfer-tanulást emellett nehezíti az is, ha a céldoménben kvalitatív értelemben túlzott számú kombinációhoz hiányoznak megfigyelések.

Dolgozatomban egy olyan kevert kvalitatív-kvantitatív transzfer-tanítási megközelítést javaslok és alkalmazok az adott problémára, melyben a tanulás tárgyát képezi egy folytonos transzfer-modell és egy kvalitatív transzfer-modell is. A kvalitatív következtetés ismert módszereinek segítségével utóbbi lehetővé teszi, hogy a kvalitatív átviteli modell tanulása során szakértői szabályokat fogalmazzunk meg és kényszerítsünk ki, valamint, hogy a két domén közötti minőségi, illetve szabályszerűségi különbségeket szakértő által interpretálható módon vizsgáljuk. A kvalitatív-kvantitatív modellpár alkalmazása többféleképpen is történhet, például oly módon, hogy a kvalitatív modell a céldomén fölötti predikció hihetőség vizsgálatát támogatja.

A javasolt és prototipizált módszereket egy nagyméretű Egyesült királyságbeli adatkészletből egy konkrét ipari magyarországi kisméretű adatkészletre való leképezés problémáján demonstrálom.

Abstract

It is extremely important to ensure energy efficiency nowadays, both from a financial and climate protection point of view. A significant part of the energy consumption is the operation of the buildings, therefore it is important to optimize the energy consumption of buildings. The main step in the optimization process is to determine the energy efficiency of the building and the corresponding expected consumption, which serves as a benchmark when considering the actual consumption.

The energy efficiency of buildings is usually characterized by energy certificates, which include, among other things, the energy characteristics of the building and the expected consumption of the building. However, many buildings are not certified, and the certification process requires complex calculations and sometimes on-site visits.

The general goal of my research is to replace the human effort required during energy certification with a statistical approach, so that the estimation of expected energy consumption can be generalized and scaled. The statistical model is based on the qualitative energetic characteristics of the buildings (e.g. wall, heating system quality) and the related expected consumption. The quality characteristics and the expected consumption value can be extracted from the energy certificates - with appropriate conversions. When determining the expected value, you can determine the quality characteristics with an energetic visual inspection, so there is no need for detailed assessment and calculation.

One of the main challenges of the research is that while some countries have open and extensive certificate databases, other countries do not publish data of sufficient quantity or quality. The focus of my research is the use of transfer learning for the statistical determination of expected consumption, even in countries where little or incomplete data is available. The transfer of knowledge is not a trivial problem, since not only the distribution of buildings varies from country to country, but different methods are used to calculate the expected values. During transfer learning, if the source and target domains are logically significantly different from each other, negative transfer can also occur.

In my thesis, I propose and apply a mixed qualitative-quantitative transfer teaching approach to the given problem, in which the subject of learning is both a continuous transfer model and a qualitative transfer model. With the help of the known methods of qualitative inference, the latter makes it possible to formulate and enforce expert rules during the learning of the qualitative transfer model, as well as to examine the differences in quality and regularity between the two domains in a way that can be interpreted by an expert. The qualitative-quantitative model pair can be applied in several ways, for example, in such a way that the qualitative model supports the examination of the reliability of the prediction over the target domain.

I present the proposed and prototype methods for the problem of mapping between the large data set in the United Kingdom and the specific small data set in industrial Hungary.

1. fejezet

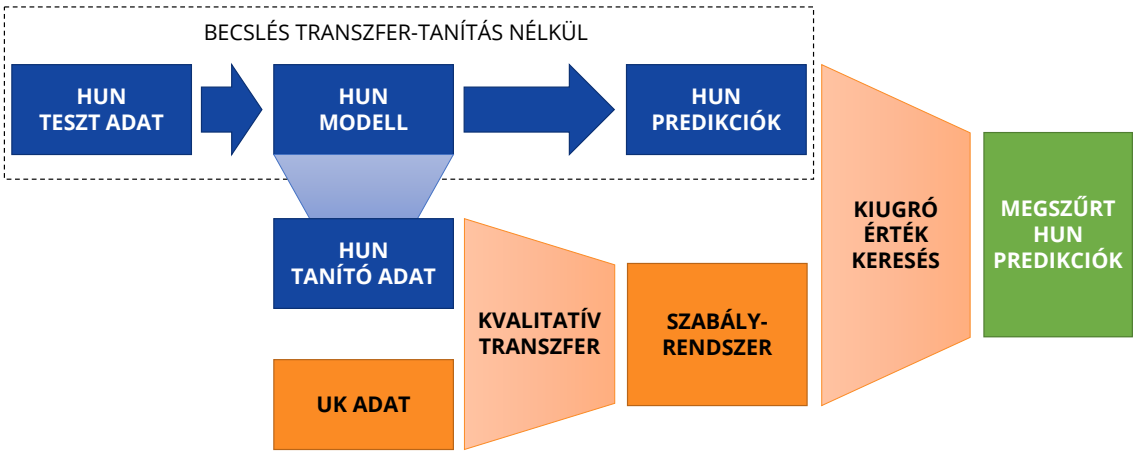
Bevezetés

Az energiahatékonyság napjainkban egyre fontosabb kérdés, az energiaárak rohamtempóban nőnek, és egyre jobban érezzük a túlfogyasztás okozta klímaváltozást. Külön figyelmet kell szentelni az épületekre, mivel energiafogyasztásuk olyan mértékű terhelést jelent a környezetnek, mint a közlekedési vagy az ipari szektor [17].

Az épületek fogyasztás-csökkentési potenciáljának meghatározásakor tisztában kell lennünk az épület energiahatékonysági jellemzőivel. Egy jó visszamérési metrikaként szolgálnak az energetikai tanúsítványok, azonban nem minden épület rendelkezik velük. Egy új tanúsítvány (és ezzel együtt egy elvárt fogyasztás referenciaérték) meghatározása igen költséges lehet, így törekednünk kell az egyszerűbb, mégis elfogadható megoldások kidolgozására. Munkám célja, hogy gyorsan és olcsón megfelelő pontosságú becslést tudjunk adni – tanúsítvánnyal nem rendelkező – ingatlanok elvárt fűtésfogyasztására.

Egy, a leíró kategoriák értelmében terjedelmes tanúsítványkészlet jó alapként szolgálhat egy elvárt-fogyasztás becslésére alkalmas modell felépítésére. Dolgozatom fő motivációja, hogy Magyarország nem rendelkezik kiterjedt, publikusan elérhető tanúsítványadatbázissal. A feladatot céges együttműködés keretében valósítottam meg, így rendelkezésemre állt egy pár ezres magyar energetikai tanúsítvány adatkészlet, azonban ez nem fedi le reprezentatívan a teljes magyar ingatlanállomány minden szegmensét.

A munkám alapgondolata, hogy publikus – külföldi országok – tanúsítvány adatbázisokból átemeljük a tudást – különböző korrekciók által – hazai környezetbe. Ennek során azzal az alpfelvetéssel élek, hogy a tanúsítványokat jellemző fő összefüggések megegyeznek két különböző országot vizsgálva.



1.1. ábra. Kiugró érték keresés kvalitatív-kvantitatív transzfer tanítás segítségével

A dolgozatban bemutatott munka áttekintése a 1.1 ábrán látható. A munkám egy meglévő magyar fogyasztás-becslő modell felülvezérlését szolgája egy – kvalitatív transzfer-tanítás által előállított – szabályrendszer által. A transzfer-tanítás során egy Egyesült Királyságbeli adatkészletet vettem alapul, amely több millió tanúsítvánnyal rendelkezik.

A munkám során egy kvalitatív transzfer-tanítást valósítottam meg, melynek lényege, hogy egy közös kvalitatív modellt állítok elő a két országra nézve. A közös leképezésnek köszönhetően a szabályszerűségek átültethetőek a két domén között. A kvalitatív transzfer-tanítással felépített szabályrendszert a magyar modell kiugró értékeinek ellenőrzésre használom. A megközelítést valós adatokon teszteltem, és az általam kitűzött célnak megfelelő eredményeket kaptam.

A kvalitatív modellezés más megközelítések alkalmazását is lehetővé teszi, például bevezethetőek különböző szakértői (mérnöki) szabályok az egyes épületcsoportok tekintetében. Ezenkívül számos lehetőséget kínál még a kvalitatív következtetési eljárások használata.

Munkám fő – tudományos értelemben – vett újdonsága, hogy a transzfer-tanítás során a kvalitatív megközelítések alkalmazása nem számít bevett gyakorlatnak.

2. fejezet

Háttérismeretek

A jelen fejezet ismerteti az általam végzett kutatás megértéséhez szükséges háttérismeretket. Elsőként az energetikai tanúsítványokat, és az épületek főbb energetikai jellemzőit mutatom be röviden. Ezt követően az energetikai tanúsítvány adatbázisok helyzetét ismertetem elérhetőség szempontjából, majd ismertetem a rendelkezésemre álló magyar, illetve potenciális külföldi adatbázisokat. Az utolsó alfejezetben ismertetem a transzfer-tanulás módszerét és fajtáit.

2.1. Energetikai tanúsítványok

Az energetikai tanúsítványok szabályozása országonként eltér, azonban a céljuk közös: átfogó képet adjon ingatlanunk energetikai állapotáról, és energiaigény csökkentő javaslatokat tegyen. Ezenkívül tanúsítvány alapján egy adott épület energetikai minősége összevethető a követelményekkel, épületek energetikai minősége összehasonlítható és energetikai besorolás alá vethető. A továbbiakban a magyar energetikai tanúsítványokról írok részletesebben, azonban – főleg az EU tagországokban – nagyon hasonlóak.

Meglétük sok esetben kötelezettség is, például új épület építése vagy energetikai pályázatok igénybevétele esetén. Kibocsátásuk engedélyhez kötött, és tartalmi követelményrendszerük törvényileg szabályozott.

Elkészítésük során a tanúsítónak rendelkezni kell az épület szükséges energetikai paramétereivel. Az energetikai jellemzőket a megrendelő biztosíthatja műszaki dokumentumok formájában, vagy kérhet helyszíni felmérést, Amely nyilván erőforrásigényesebb.

A tanúsítvány melléke egy javaslat is, ami az épület energiaigényének csökkentésére vonatkozik. A tanúsítvány tartalmazza azt is, hogyan változik az épület energiaigénye a javaslat megvalósítása után.

Az épületek energetikai tanúsítása a 2006 óta érvényben lévő energetikai és hőtechnikai szabályozás szerint történik. Az energetikai méretezését és megfelelőségét a 7/2006TNM [16] rendelet 1. melléklete alapján kell véghezvinni.

A követelményértékek azonban 2018-ban és 2021-ben is változtak. Ez azért releváns, mert a követelményértékek és az építés év alapján alsó becslést lehet adni az energetikai jellemzőkre.

2.1.1. Épületek energetikai jellemzői

Munkám során energetikai tanúsítványokból nyerem ki az energetikai jellemzőket és fűtésfogyasztás értékét. Ebben a fejezetben a tanúsítványokba foglalt főbb energetikai jellemzőket mutatom be áttekintő módon. Az energetikai tanúsítványok többek között a következő energetikai jellemzőkre térnek ki [15]:

- a külső határoló szerkezetek és nyílászárók méretei, hőtechnikai jellemzői,
- az épületben található világítóberendezések fajtái, illetve esetleges vezérlése,
- a kazán fajtája, a fűtési alapvezeték hossza, elhelyezkedése, hőszigetelése,
- a fűtésrendszer vezérlése vagy szabályozási eljárása,
- a HMV (használati melegvíz) termelésének (központi, egyedi, átfolyós, tárolós) és hálózatának (víztakarékos szerelvények, egyedi mérés) jellemzés
- klimatizáló és légtechnikai berendezések megléte és minősége
- valamint a használt energiahordozók.

2.1.1.1. Épületszerkezetek energetikai jellemzői

Az épületek legfontosabb energetikai tényezője a épületszerkezet energetikai minősége. A számítások végén az egyes szerkezeti elemekhez U értéket kell rendelni (a munkám során ezeket a folytonos U értékeket kategorizáltam. Az U érték definíció szerint:

Hőátbocsátási tényező, azt a hőmennyiséget határozza meg, amely az adott szerkezet 1 m^2 felületén 1 másodperc alatt átáramlik, amikor a külső és belső hőmérséklet különbsége 1 fok. Az alacsonyabb érték kedvezőbb hőtechnikai jellemzőt jelent [28].

Az épületszerkezet jellemzésekor a következő tényezőket kell számításba venni: külső falak, homlokzati nyílászáró szerkezetek, lapostetők, beépített tetőtereket határoló szerkezetek, padlásfödémek, pincefödémek és árkádfödémek [15] .

Ezen szerkezeti elemeknél energetikai szempontból a legfontosabb, hogy milyen anyagból készültek (van-e szigetelés), milyen vastagok és geometriai elhelyezkedésük, fizikailag mivel határosak.

2.1.1.2. Épületgépészet energetikai jellemzői

Épületgépészeti szempontból az épületek hűtő-fűtő, valamint használati melegvíz rendszereit kell megvizsgálnunk. A vizsgálat többek között a következő pontokra tér ki.

Meg kell határozni a fogyasztói kört, épületrészt, ezenfelül az üzemeltetési szokásokat, üzemidőt. Fel kell mérni a hőtermelők, illetve szabályzók műszaki állapotát. Meg kell határozni a rendszerelemek (hőleadók, ventilátorok, melegvíz termelők, tárolók, szivattyúk, hűtőberendezések, vezetékrendszerek) fontosabb tulajdonságait, műszaki állapotát [15].

2.2. Energetikai tanúsítványok adatbázisai

Munkám során energetikai tanúsítványokat használok a becslő modellek tanítására. Magyarország nem publikált nyilvánosan elérhető tanúsítvány-adatbázist, azonban céges forrásból rendelkezem korlátozott számú magyar adattal.

A transzfer tanulás megvalósításához külföldi, nyilvános adatbázisokat lehet használni. Jelen fejezetben ismertetem a rendelkezésemre álló magyar, illetve lehetséges külföldi adatbázisokat.

Az adatbázisok kiválasztása során azt is meg kell vizsgálni, hogy az adott adatbázis rendelkezik-e a megfelelő attribútumokkal. A mi esetünkben a következő fizikai objektumok kvalitatív jellemzése szükséges: fal, ablak, tető, padló, fűtés, fűtésvezérlés és melegvízelőállítás. Ezen kívül az adatbázisnak rendelkeznie kell néhány folytonos változóval leírható alapadattal, amelyek a következők: építés éve, alapterület, belmagasság és fűtésfogyasztás.

A feladatnak megfelelően az adatbázisokra az épület jellegének megfelelően is rászűrtem, azaz eldobtam a lakás típusú ingatlanokat.

2.2.1. Magyar adatbázis

Publikus magyar adatbázis híján egy magyar adatbázis előállítása is a munkám része volt. A rendelkezésemre álló – XLSM (makróbarát XLSX) formátumú – tanúsítványokat automatizálva dolgoztam fel, és egy adatbázisba rendeztem őket. Az adatbázis olyan épületeket tartalmaz, amelyek 2012-től 2022-ig lettek energetikailag felmérve. A feldolgozást egy Python script végzi, ami számos tisztítási és kvantálási lépést valósít meg. A kvantálás során az egyes energetikai jellemzőket a 2.2.2 alfejezetben bemutatott adatbázis értéktartományai szerint végeztem. A feldolgozás után a magyar adatbázis nagyjából 2500 tanúsítványt tartalmaz, ami körülbelül 4000 nyers adat feldolgozásából állt elő.

2.2.1.1. Adatok feldolgozása

Az alapadatok kinyerése könnyű folyamat: az építés éve, a belmagasság és a fűtésfogyasztás átmásolható egyszerűen. A fűtésfogyasztást az egész munkám során légköbméterre vetítve vizsgálom. Az teljes alapterületet – amennyiben több emeletes épület – pedig megkapjuk az összes emelet alapterületének összeadásával.

A kvalitatív adatok kinyerése néhol már nagyobb kihívást jelentett. A falazat, ablakok és padlózat U értéke egyszerűen átmásolható.

A tetőszerkezet U értékét a 2.1 egyenlet alapján kapjuk meg, amennyiben a következő jelöléseket alkalmazzuk: U_{lt} (lapostető U érték), A_{lt} (lapostető terület), U_{mt} (magastető U érték), A_{mt}, U_{p} (pincefödém U érték), A_{p} (pincefödém terület), $A_{\mathrm{öt}}$ (összes tető terület).

$$\frac{U_{\mathrm{lt}} * A_{\mathrm{lt}} + U_{\mathrm{mt}} * A_{\mathrm{mt}} + U_{\mathrm{p}} * A_{\mathrm{p}}}{A_{\mathrm{öt}}} \tag{2.1}$$

Az kvalitatív épületszerkezeti adatokon kívül a gépészeti jellemzőket is kinyertem a nyers adatokból. A fűtőrendszer energiahatékonyságát a 2.2 képlet írja le a következő jelöléseket használva: α (egységnyi kibocsátás tényező), β (hőtermelő teljesítménytényezője), γ (reszponzívitási tényező). Az egységnyi kibocsátás tényezőt (gáz esetén mindig 0.222) meg kell szorozni a hőtermelő teljesítménytényezőjével, ezenkívül egy reszponzívitási tényezővel. A reszponzivítási tényezőt a 2.3 képlet szerint kapjuk meg, a következő jelöléseket használva: δ (elosztóvezeték fajlagos vesztesége), η (fűtés éves nettó hőenergia igénye).

$$\alpha * \beta * \gamma \tag{2.2}$$

$$\frac{1+\delta}{\eta} \tag{2.3}$$

A melegvíz-előállítás energiahatékonyságát a fűtőrendszer energiahatékonyságának megfelelően számítjuk ki. Amennyiben ugyanaz a rendszer látja el mindkét feladatot, akkor a két érték megegyező lesz. Ha a melegvíz előállítás villamos energia segítségével történik (pl. bojler), akkor az egységnyi kibocsátás tényező 0.381.

Következő lépésként a kinyert folytonos energetikai jellemzőket kvantáltam. Ez a diszkretizáció a 2.2.2 alfejezetben található adatbázis értéktartományai szerint végeztem. Az épületszerkezeti tulajdonságok kvantálása az 2.1, az épületgépészeti adatok az 2.2 táblázatban látható értéktartományok mentén történt.

	Ablak U érték	Fal U érték	Tető U érték	Padló U érték
1	≥ 4.4	≥ 1.6	≥ 1.0	≥ 0.7
2	<4.4	<1.6	<1.0	<0.7
3	<3.3	<1.01	<0.5	<0.45
4	<2.5	<0.61	<0.3	<0.3
5	<1.7	<0.3	<0.15	<0.2

2.1. táblázat. Épületszerkezeti jellemzők kvantálási értéktartományai

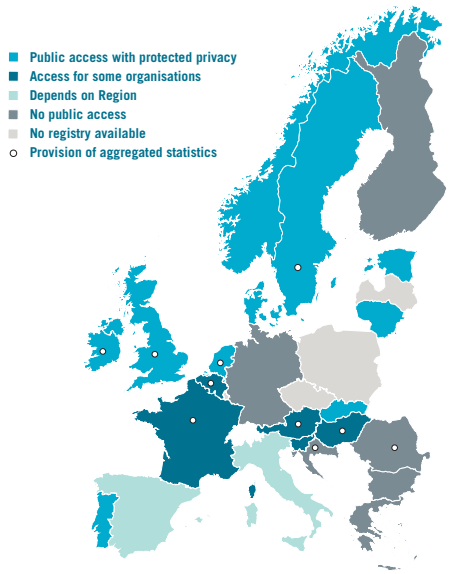
	Fűtés hatékonyság	Fűtésvezérlés hatékonyság	Melegvíz-előállítás hatékonyság
1	≥ 0.55	≥ 15	≥ 0.55
2	<0.55	<15	< 0.55
3	<0.44	<9.6	<0.44
4	<0.33	<5.5	<0.33
5	<0.22	<3.3	<0.22

2.2. táblázat. Épületgépészeti jellemzők kvantálási értéktartományai

A kvantálás után még néhány adattisztítási lépést végeztem el. Azokat az épületeket el kellett dobnom, ahol nem volt értelmezhető évszám az építés évét tekintve. Ezenkívül csak olyan épületeket tartottam meg, amelyek gázt használnak fűtésre. Az elvégzett lépések után egy CSV fájlba exportáltam az adatbázist.

2.2.2. Külföldi adatbázisok

A magyar adatbázis tudás-bővítésére – a javasolt megoldásom szerint – külföldi energetikai tanúsítvány-adatbázisokat lehet használni. Minden ország más-más minőségben publikál adatokat, és azok publikus elérhetősége is változó. Az EU-s országok tanúsítvány-adatbázisainak elérhetőségét az 2.1 ábra mutatja be. A tanúsítvány-adatbázisok összevetésére léteznek részletesebb összehasonlítások is például a minőségükre vonatkozóan [20].



2.1. ábra. Energetikai tanúsítvány-adatbázisok az EU-ban elérhetőségük szerint [1]

A feladat megoldására elsőként az Anglia és Wales tartományt lefedő tanúsítvány-adatbázist választottam. Döntésemet az indokolta, hogy az említett adatbázis teljesen elérhető kutatási célra, illetve tartalmazza az összes szükséges attribútumot a munkámhoz.

A munkám további részeként újabb tanúsítvány-adatbázisokat fogok megvizsgálni, és használni transzfer tanulás céljából.

2.2.2.1. Anglia és Wales

Anglia és Wales tartományban 2008-tól a lakossági épületekhez is kötelező csatolni a tanúsítványok mögöttes számításait. A lakossági épületek esetében az adatszolgáltatás mindig kötelező követelmény volt [18]. A teljes adatbázis több mint 22 millió lakossági és több mint 1 millió nem lakossági tanúsítványt tartalmaz. A munkám során a lakossági adatokat használtam, mivel csak a lakossági tanúsítványok rendelkeznek a megfelelő energetikai jellemző leírókkal. Az Anglia és Wales tanúsítványait tartalmazó adatkészletre továbbiakban UK adatbázisként hivatkozom.

Az adatok feldolgozása során egyszerűbb dolgom volt, mint a magyar adatok esetében, mivel a tanúsítványok már táblázatba voltak foglalva városonként. A feldolgozáshoz ebben az esetben is egy Python scriptet használtam, ami összegyűjtötte az összes város tanúsítvány-adatbázisát, és kinyerte belőlük a szükséges adatokat.

Az adatkinyerés során kiszűrtem azokat az adatokat, amelyek nem rendelkeztek alapterülettel, építés évével, fűtőanyag típussal vagy fogyasztás költség értékével. Ezután eldobtam azokat a sorokat, ahol nem gáz az energiatípus. Az adatbázisban csak a ház típusú ingatlanokat tartottam meg.

A következő lépés a fűtésfogyasztás értékének előállítása volt. Erre azért volt szükség, mert az adatbázis nem tartalmazta a fűtésfogyasztás értéket, csupán a fűtésfogyasztás költségét. A feladatom az volt, hogy a költségből visszafejtsem a fogyasztott mennyiséget. Ehhez összevetettem a historikus fűtésfogyasztás árakat [3] az adott tanúsítvány felmérési évével, majd az alapdíj és egységár segítségével meghatároztam minden épület fűtésfogyasztását.

Mindkét adatkészlet esetében normalizáltam a fűtésfogyasztás értékét: elosztottam az alapterülettel és a belmagassággal. Ezáltal az épületek fogyasztás szempontjából a kimeneti értékekkel összehasonlíthatóvá válnak.

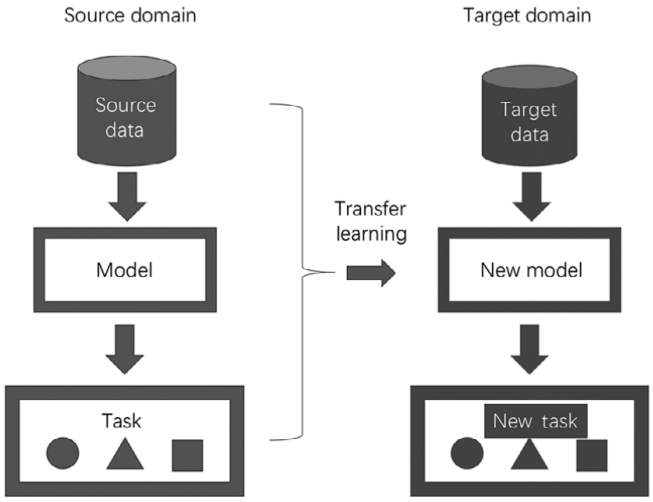
További adattisztítási lépés volt az építés év meghatározása. Az adatbázisban szövegesen építési év tartományok voltak megadva. Ezeknek a tartományoknak vettem a közepét, és hozzárendeltem az adott tanúsítványhoz.

A kvalitatív energetikai jellemzők szövegesen voltak megadva, ezekhez értelemszerűen számértékeket rendeltem (pl. very poor $\rightarrow$ 1).

Az elvégzett lépések után egy CSV fájlt állítottam elő, amely nagyjából 2,5 millió energetikai tanúsítványt tartalmazott. Az előállított formátum megegyezett a magyar adatkészlettel, így a két adatbázis teljesen összehasonlítható lett.

2.3. Transzfer-tanulás

A transzfer tanulás során az a célunk, hogy javítsunk egy új feladaton egy kapcsolódó feladatból származó tudás átadásán keresztül [25]. A transzfer tanulás legtöbbször képfeldolgozás során jön elő, amikor egy modell előre be van tanítva egy alap adatkészleten, majd egy speciális feladatra a modellt finomra hangolják. Azonban, ahogy a dolgozatom témája is mutatja, más területeken is előkerülhet a transzfer tanulás. A szakirodalom is megemlít transzfer tanulás alkalmazást hasonló jellegű városi számítástechnikai problémában [26]. Egy ilyen klasszifikációs probléma például a sanghaji levegőminőség előrejelzése pekingi forrásadatok alapján.



2.2. ábra. Transzfer tanulás folyamata [26]

A 2.2 ábra egy transzfer-tanulási folyamatot szemléltet, ahol a célterület hiányos adatkészletét a forrástartományból vett tudással kompenzáljuk. Az ábra bal oldala egy klasszikus gépi tanulási folyamatot mutat be, a jobb oldalon egy transzfer-tanulási folyamat látható. Egy transzfer-tanulási folyamat nem csak a céldomén adatait használja fel, hanem a forrás tartomány bármely részét (tanító adatot, modellt, feladatot) is használhatja.

Transzfer tanulás során nem csak arra kell ügyelnünk, hogy hogyan valósítsuk meg, hanem arra is, hogy mikor alkalmazzuk. Ahol a forrás- és célterület nem kapcsolódik kellően egymáshoz, negatív transzfer alakulhat ki, ami rosszabb lehet, mintha nem alkalmaznánk semmilyen transzfert. Fontos tehát az adatkészletek megfelelő összehasonlítása mielőtt transzfer tanulást indítunk.

2.3.1. Transzfer-tanulás motivációja

A transzfer-tanulást többek között a következő problémák indokolják. Általánosságban a gépi tanuláshoz nagy mennyiségű adatra van szükségünk, azonban számos célterületen nem áll rendelkezésre elegendő adat, vagy nagyon költséges lenne az adat "előállítása" (pl. felcímkézése). Kis mennyiségű adatnál nagyon magas a túltanulás kockázata is. A gépi tanulással létrehozott modelleknek általában robosztusnak is kell lenniük. Alapfeltételezésként élünk legtöbbször azzal, hogy a tanító és teszt adatkészletek eloszlása megegyezik, azonban ez időről időre változhat. Egy robusztus modell képes lekövetni ilyen változásokat.

Számos olyan esettel is találkozhatunk, amikor személyre szeretnénk szabni egy adott szolgáltatást, azonban – például adatvédelmi okokból – nem rendelkezünk a célszemélyről elég információval. Ilyenkor az egyedi kevés információt egy alapmodellre szeretnénk ültetni, így a közös tudás alkalmazható az adott egyénre is. Ugyanígy léteznek olyan esetek is, amikor maga az adat nem osztható meg cégek között, viszont magát a tudáscserét meg szeretnénk valósítani például egy közös modell építése érdekében.

2.3.2. Transzfer-tanulás definíciója

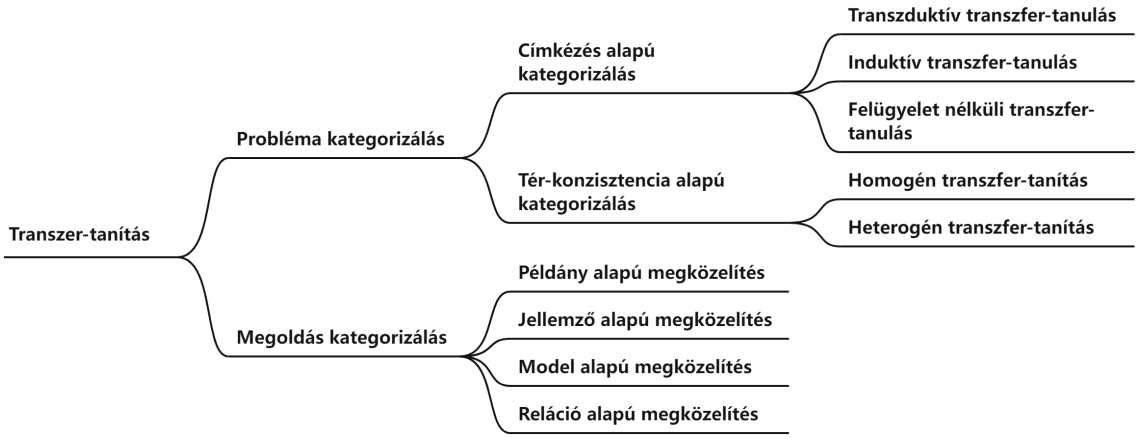
A jelen alfejezetben a transzfer-tanulás definícióját ismertetem. Egy $\mathbb{D}$ domén két komponensből áll: egy $\mathbb{X}$ jellemzőtérből és egy $\mathbb{P}_x$ ($\mathbb{x}$: input adat) peremeloszlásból, ahol minden $\mathbb{x} \in \mathbb{X}$. Ez alapján egy domént a következőképpen tudunk felírni: $\mathbb{D} = \{\mathbb{X}, \mathbb{P}_\mathbb{x}\}$. Ha két különböző domént vizsgálunk, akkor azok a jellemzőteret vagy a peremeloszlást nézve kü-

lönböznek egymástól. Egy $\mathbb{T}$ feladat két komponensből áll: egy $\mathbb{Y}$ címketérből és egy $ft(\cdot)$ függvényből. Egy feladatot a következőképpen írhatunk fel: $\mathbb{T} = \{\mathbb{Y}, (ft(\cdot))\}$. Egy osztályozási feladatot tekintve az címkék értékei ($\mathbb{Y}$) lehetnek binárisak, vagy diszkrét értékűek többosztályos feladatot nézve. Egy regressziós problémát tekintve az a címkék folytonos értékűek [26]. A bemutatott jelöléseket használva egy transzfer-tanulási probléma a következőképpen definiálható.

Definíció 1 (transzfer-tanulás). Adott egy forrásdomén $\mathbb{D}_s$, egy tanulási feladat $\mathbb{T}_s$, egy céldomén $\mathbb{D}_t$ és egy tanítási feladat $\mathbb{T}_t$. Az transzfer-tanulás célja, hogy segítse az $ft(\cdot)$ prediktív függvény tanulását a céltartományban a $\mathbb{D}_s$ és $\mathbb{T}_s$ tudás segítségével, ahol $\mathbb{D}_s \neq \mathbb{D}_t$ vagy $\mathbb{T}_s \neq \mathbb{T}_t$ [26]. ▪

2.3.3. Transzfer-tanulás kategorizálása

A transzfer-tanulást többféleképpen is lehet kategorizálni, a lehetséges kategorizálási irányokat a 2.3 ábra mutatja be. Kategorizálhatjuk a megoldandó problémát az adatok címkézettsége, vagy a két domén összefüggősége szerint. A transzfer-tanulási feladatokat csoportosíthatjuk a megoldás módja szerint is.



2.3. ábra. Transzfer tanulás lehetséges kategorizálásai [27]

2.3.3.1. Kategorizálás a probléma szerint

Az egyik kategorizálás címkézés szempontjából vizsgálja a feladatokat. Amennyiben csak a forrásdomén rendelkezik címkékkel, transzduktív transzfer-tanulásról beszélünk. Azokban az esetekben, amikor a céldomén rendelkezik címkékkel, induktív transzfer-tanulás a probléma. Szélsőséges esetekben, amikor az egyik domén se rendelkezik címkézett adattal, akkor felügyelet nélküli transzfer-tanulással jellemezhető a feladat.

Egy másik kategorizálás alapján azt vizsgáljuk, hogy mekkora konzisztencia a forrás, és céldomén között. Amennyiben $\mathbb{X}_s = \mathbb{X}_t$ és $\mathbb{Y}_s = \mathbb{Y}_t$, akkor homogén transzfer-tanulásról beszélünk. Ellenkező esetben, ha $\mathbb{X}_s \neq \mathbb{X}_t$ vagy $\mathbb{Y}_s \neq \mathbb{Y}_t$, akkor heterogén transzfer-tanulás a probléma [27].

A dolgozatban vizsgált probléma egy induktív homogén transzfer-tanulási problémának tekinthető, mivel mindkét adatkészlet címkézett adatokkal rendelkezik, ahol a jellemzőtér és a címketér megegyező.

2.3.3.2. Kategorizálás a megoldás szerint

A példány alapú megközelítések alkalmazásakor általában az a motiváció, hogy a forrásdomén címkézett adatait nem lehet közvetlenül felhasználni (a tartománybeli eltérések

miatt), viszont egy részüket megfelelően mintavételezve, vagy újrasúlyozással fel tudjuk használni a céltartományban.

Egy magasabb szintű megközelítés, hogy már egy meglévő, betanított modell szintjén (model-based) visszük át a tudást. Ilyenkor például egy apriori modell segítségével javítunk a kevés darabszámú célterület feladatmegoldásában.

A jellemző alapú megközelítések mögött az az elképzelés, hogy megtanítunk egy – mindkét domén számára – megfelelő jellemző reprezentációt mind a forrás, mind a céltartományt tekintve. Amennyiben leképezzük a forrástartomány adatait az új reprezentációra, a céltartomány tanítása során az adatok felhasználhatóvá válnak. A transzfer-tanulást ebben az esetben a közös jellemző reprezentáció jelenti [26].

A reláció alapú megközelítést elsősorban relációs problématerületeknél alkalmazzák, ahol a problémák felírhatók relációkkal, például hálózatokkal. Az ilyen típusú megoldások esetén a forrásdoménben megtanult szabályok visszük át a forrástartományba. A reláció alapú transzfer-tanítással relatív kevés tanulmány foglalkozik, ezért az itt végzett kutatások altalánosan véve tudományosan úttörő munkáknak számítanak [27].

Egy reláció alapú transzfer-tanítást megvalósító megoldás a *TAMAR* [14] algoritmus, amely a relációs tudást egy Markov-logikai hálózat [22] segítségével ülteti át. A megoldás elsőként egy predikátum leképezést készít a két tartomány között, majd egy finomítást végez felülvizsgálatok és újrasúlyozások által a megvalósított leképezéseken.

2.3.3.3. Transzfer tanulás alkalmazása a szakterületen

A dolgozatomban bemutatott probléma szintén egy relációs problémának tekinthető. A folytonos változók kvantálása után minden épület leírható kategorikus változók kombinációjaként. A kimeneti változó (fűtésfogyasztás) kvantálásával tulajdonképpen egy relációt kapunk a bemeneti kategóriakombinációt és a kimeneti kategóriát tekintve. Miután mindkét adatkészletből (magyar, UK) kinyertük a relációkat, megkezdődhet a reláció alapú transzfer tanítás. A kutatásom célja, hogy a UK doménben fellelhető összefüggéseket átvigyem a magyar tartományba, és a tudást felülvizsgálati célokra használjuk.

Kutatásom azért tekinthető innovatívnak, mivel nem csak az energetikai szakterületen, de más doménekben sem tekinthető bevett eljárásnak a transzfer-tanulás által létrehozott kvalitatív szabályrendszerek használata.

3. fejezet

Kapcsolódó kutatások

Az energiafogyasztás csökkentésének lehetséges módjait széleskörűen kutatják. Sok kutatást találhatunk az épületek fogyasztási karakterisztikájának meghatározására, *benchmarking* módszerek alkalmazására. A fogyasztási teljesítményértékelés legtöbbször idősoros adatok segítségével történik. Ezenkívül léteznek megoldások, ahol energetikai jellemzők (pl. energetikai tanúsítványok) alapján becsülnek fogyasztási értéket.

3.1. Energetikai hatékonyság becslése Walesben

A munkám során használt Egyesült Királyságból származó adatkészletet 2.2.2 más kutatásokban is használják. Ez annak köszönhető, hogy az említett adatkészlet rendkívül részletes és terjedelmes. A következő kutatás [6] is a UK adatkészletet használja energetikai hatékonyság becslésére.

A tanulmány célja, hogy a hiányzó energetikai besorolásokat (Energy Efficiency Rating (EER)) kipótolják. A kutatás a szöveges energetikai leírókat veszi alapul (pl. szigetelt fal, nincs termosztát), amelyek adattisztításra szorulnak. A tanúsítvány adatokat feltárták, megtisztították és három különálló szolgáltatáskészletet hoztak belőle létre. Létrehoztak egy adatvezérelt megközelítést, ahol klasszikus statisztikai módszerekkel összegezték a szöveges leírókat. A második megközelítés során domainvezérelt módszert használtak, azaz energetikai szabályoknak megfelelően hoztak létre összevont kategóriákat az érthetőség érdekében. A harmadik megközelítés során nem módosítottak semmit az adatkészleten, csupán előkészítették a gépi tanulásra (pl. kódolták a kategorikus változókat). Az utolsó esetben rendkívül sok változó keletkezett, összesen 1030 darab, ami a tanítás hatékonyságát rontja.

A három megközelítést különböző tanítási módszerekkel tesztelték és értékelték ki. A három megközelítés közül az utolsó hozta a legjobb teljesítményt XgBoost-ot [7] alkalmazva: nagyjából 70% pontossággal (accuracy) képes az energetikai kategória besorolást eltalálni. A folytonos leíró változók közül a teljes alapterület, a belmagasság, a lakható helyiségek száma és a bővítmények száma bizonyult a legerősebb prediktív erővel rendelkező változónak.

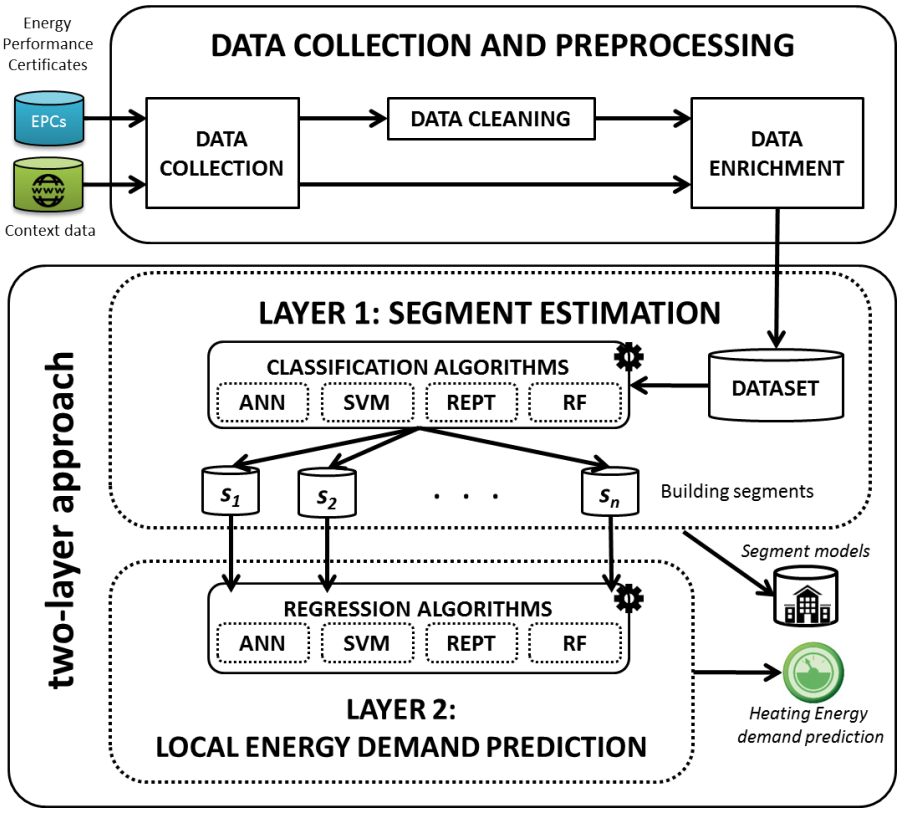
A tanulmány egy másik érdekes aspektusa arra vonatkozik, hogy hogyan lehet egy épület energetikai besorolását becsülni úgy, hogy nem állnak rendelkezésre a tanúsítványban szereplő input adatok (csupán az épület egyedi azonosítója). A kutatás erre különböző olyan proxy adatforrásokat javasol használni, amelyek a legtöbb walesi épület számára elérhetőek. A különböző proxykon keresztül többek között lekérhetőek az épületek fizikai paraméterei, például az alapterület és épület típusa. A fizikai paramétereken kívül a szomszédos épületek energetikai jellemzőit is begyűjtik. A feltételezés az, a hogy az alapvető fizikai leírók és a szomszédos épületek energetikai jellemző relatív jó kiindulási alap az ener-

getikai hatékonyság becslésére. Ez a megközelítés nagyon praktikus, hiszen nem szükséges az épületet ket felmérni, csupán le kell kérni az épülethez egyébként is hozzáférhető adatokat. Sajnos a tanulmány rámutat, hogy ez a megközelítés gyakorlatban nem használható, mivel túl gyenge predikciós képességgel rendelkezik (40% pontosság). A fő oka az, hogy ebben az esetben nincsenek konkrét energetikai leírók, csupán a szomszédos épületekre hagyatkozik a modell.

A tanulmány igen értékes olyan szempontból, hogy nagyon hasonló aspektusból vizsgálja az általam is kutatott problémát, illetve ugyanazt a UK adatkészletet használja, mint az én kutatásom.

3.2. Fogyasztásbecslés kétrétegű megközelítése

Egy komplex, energetikai tanúsítványok alapján becslő megoldást dolgoztak ki a Torinoi Politechnikumon [2]. A kidolgozott megoldásuk folyamatát a 3.1 ábra mutatja be.



3.1. ábra. Fogyasztásbecslés kétrétegű megközelítésének folyamata [2]

A megvalósított folyamat energetikai tanúsítványokat és környezeti adatokat kap bemenetként. A környezeti adatok az épület elhelyezkedésének megfelelő éghajlati viszonyokat tartalmazza. A folyamat adatgyűjtéssel és az adatok feldolgozásával kezdődik, ahol az adattisztítás után megtörténik az energetikai tanúsítványokból származó adatok kiegészítése a környezeti adatokkal.

Az előző lépésekből előáll az adatbázis, és egy kétrétegű megközelítés valósítja meg a fogyasztásbecslést. Az első réteg az épületek szegmentálását végzi. Az épületeket felosztották három csoportba az alapján, hogy az épület keveset, sokat vagy nagyon sokat fogyaszt. A felosztás után egy klasszifikációs problémát oldottak meg, amelyben a Random Forest algoritmus teljesített a legjobban (∼86% Accuracy).

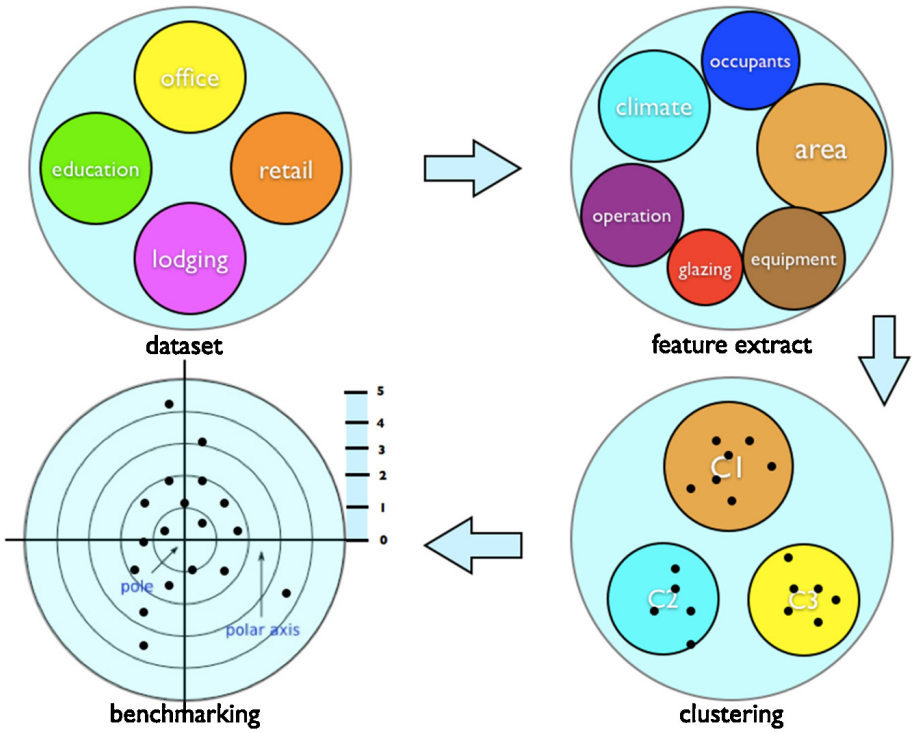
A második rétegben mindhárom épületszegmensre egy külön regressziós becslő modellt építettek többféle modellen tesztelve. A regressziós becslésre a REPT (Reduced Error Pruning Tree) algoritmust választották, mert a három szegmens hibáját átlagolva ez teljesített a legjobban (∼17% MAPE (Mean Average Percentage Error)).

A kétrétegű megközelítést validálására egy egyrétegű regressziót is kipróbáltak. Az egyrétegű megközelítés – amelyben nincs épületszegmentáció – ∼5%-al nagyobb hibát (MAPE) eredményezett a kétrétegűhöz képest.

3.3. Energetikai hatékonyság jellemzés klaszterezéssel

A jelen alfejezet egy klaszterezés alapú módszert mutat be épületek energetikai hatékonyságának jellemzésére, melyet a Pennsylvaniai Egyetemen dolgoztak ki. A *benchmarking* folyamatot az 3.2 ábra mutatja be.

Az algoritmus egyik fő különbsége az előzőhez képest, hogy nem számított elvárt értéket vesz alapul (nem tanúsítványból dolgozik), hanem mért adatokat használ. Ezzel a módszerrel egy épület összehasonlítható a hozzá hasonló tulajdonságú épületekkel. Például egy kórház a többi kórházhoz képest jól van-e üzemeltetve.



3.2. ábra. Energetikai hatékonyság jellemzés klaszterezéssel [9]

A megoldás során több adatbázist használnak, amelyek különböző funkciójú épületeket tartalmaznak. Összesen 5215 Egyesült Államokban található épületet vizsgáltak.

Az első lépésben az adatbázisokból kinyerik az energetikai szempontból fontos adatokat (terület, üvegezés aránya, klíma, üzemeltetés, stb.). Ezt követően egy klaszterező algoritmus segítségével épületcsoportokat hoznak létre, amelyek a hasonló jellemzőkkel bíró épületeket tartalmazzák. A klaszterezés befejezése után elvégezhető a *benchmarking*. A megegyező klaszterben lévő épületek összehasonlíthatóvá válnak a klaszter súlypontja szerint. A klaszter súlypontja egy ténylegesen nem létező, pszeudo-épület lesz.

4. fejezet

Feltáró adatelemzés

A munkámat részletes feltáró adatelemzéssel kezdtem, hogy megértsem a magyar és a UK adatkészlet főbb jellemzőit, illetve összehasonlítsam a két adatbázist. Az összehasonlítás fő célja, megvizsgáljam egy esetleges transzfer tanulás lehetőségeit. Ha a két adatkészlet eloszlásai és – főként – korrelációi lényegesen nem térnek el egymástól, akkor jó eséllyel megtörténhet a (reláció alapú) tudásátvitel.

A feltáró adatelemzésben először megvizsgáltam és összehasonlítottam a két adatbázis változóit. Második lépésként a korrelációkat és a két adatbázis korrelációi közötti különbséget vizsgáltam meg. Ezután főkomponens analízist végeztem.

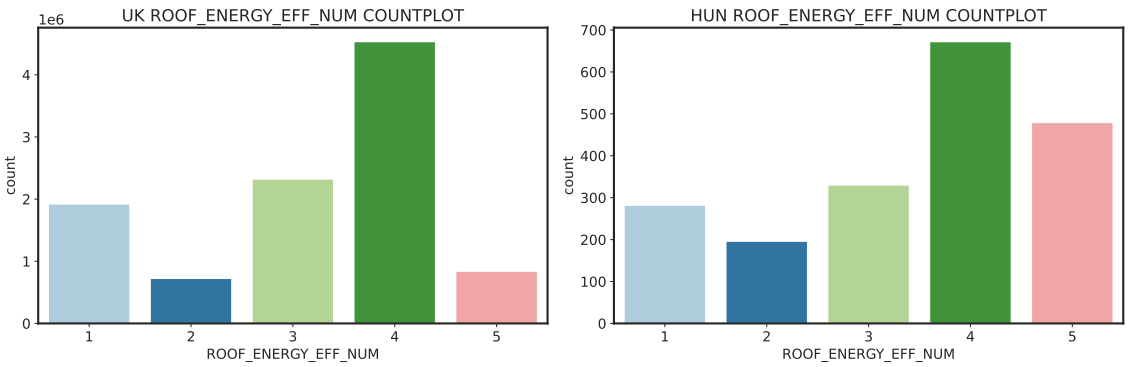
Az feltáró adatelemzés során a különböző adattisztítási és adatredukciós folyamatok után a UK adatbázis nagyjából 2,4 millió, a magyar adatbázis kb. 2500 megfigyelésből áll. A munkám további részében is ezeket az adatokat vettem kiindulási adatként.

4.1. Eloszlások összehasonlítása

A feltáró adatelemzést az eloszlások összehasonlításával kezdtem. Az ábrákon baloldalt mindig a UK, jobboldalt a magyar adatbázis eloszlási láthatóak. A kategorikus változók esetében oszlopdiagramot, a folytonos változók esetén hisztogramot alkalmaztam az eloszlások vizualizálására.

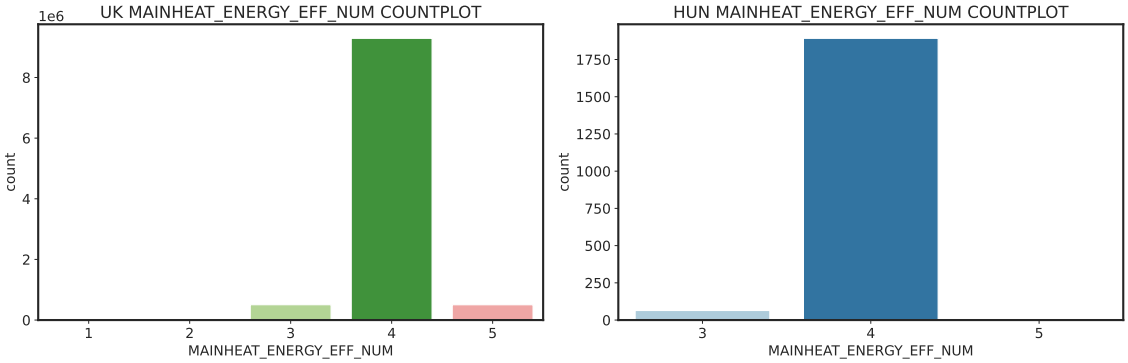
4.1.1. Kvalitatív jellemzők eloszlásai

A kvalitatív energetikai jellemzők közül a tető (4.1), fűtés (4.2), fűtésvezérlés (4.3) és melegvíz-előállítás (4.4) nagyon hasonló eloszlásokkal rendelkezik a magyar és UK adatbázist tekintve.

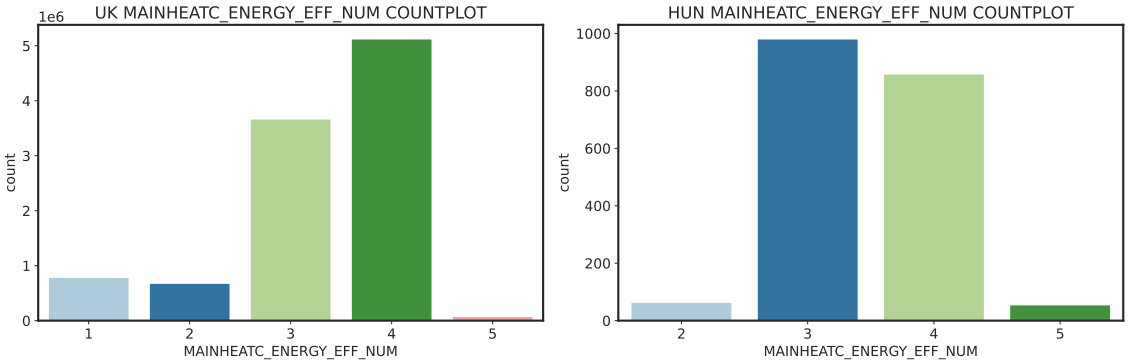


4.1. ábra. A tetőt leíró jellemző eloszlásának összehasonlítása a két adatbázisban

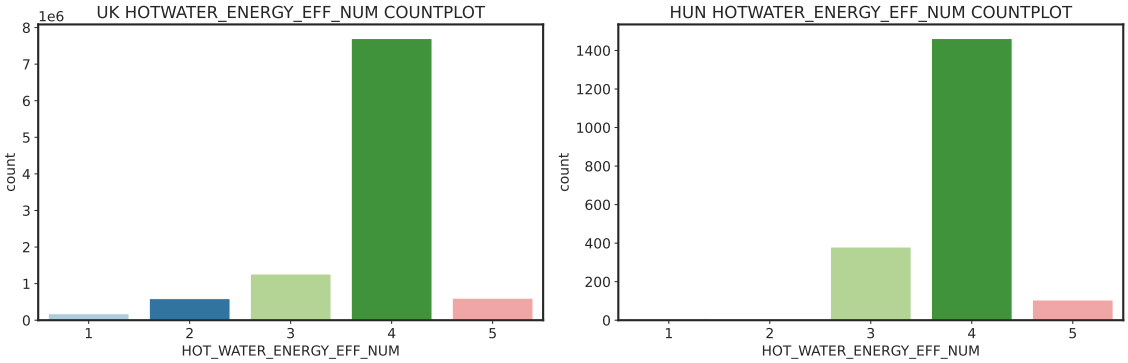
A magyar adatoknál a gépészeti jellemzőket tekintve kevés, vagy egyáltalán nincsenek az alacsony (1-es, 2-es) kategóriákban értékek. Egy épület sincs, ahol a fűtés jellemző 1-es vagy 2-es kategóriát kapott volna. Ugyanígy fűtésvezérlés változónál, ahol az 1-esek hiányoznak. A melegvíz-előállítás jellemző 1-es, 2-es kategóriájában csupán pár darab érték található. A jelenséget a UK kvantálási skálája okozza, látható mindkét adatbázis eloszlásán, hogy a kategóriákat jobban fel lehetett volna osztani.



4.2. ábra. A fűtést leíró jellemző eloszlásának összehasonlítása a két adatbázisban



4.3. ábra. A fűtésvezérlés leíró jellemző eloszlásának összehasonlítása a két adatbázisban

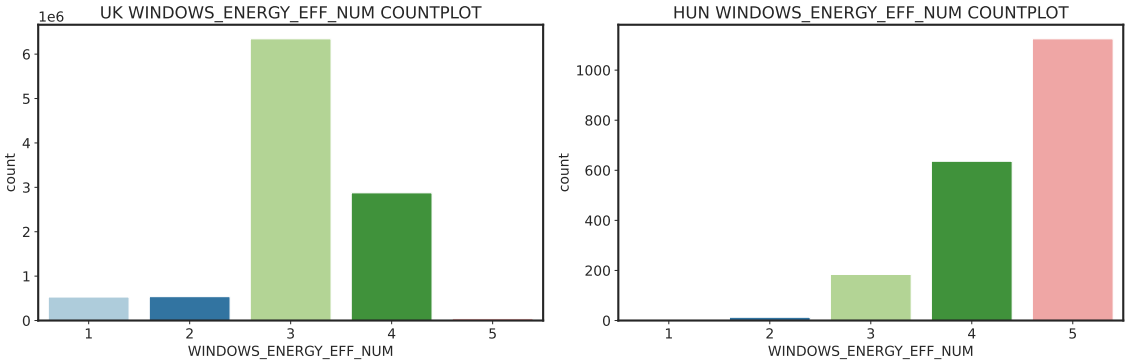


4.4. ábra. A melegvíz-előállítást leíró jellemző eloszlásának összehasonlítása a két adatbázisban

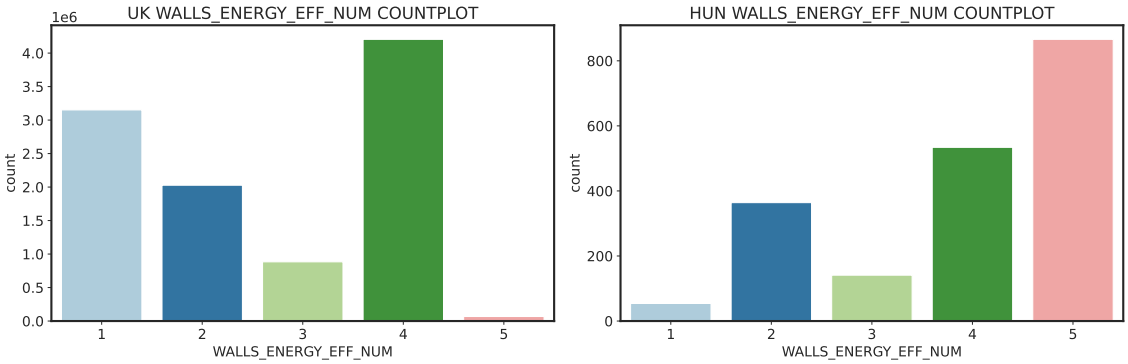
Az ablakot (4.5), falat (4.6) és padlózatot (4.7) leíró változók eloszlásainál már nagyobb eltérések fedezhetőek fel. A legfeltűnőbb jelenség, hogy a fal és az ablak esetében

nagyon kevés 5-ös érték van (a tető esetében is megfigyelhető ez, de kevésbé szembetűnően). Ezt szintén a UK adatok nem megfelelően megválasztott skálázási értéktartományai okozzák.

Ezenkívül megfigyelhető, hogy a magyar adatok tekintetében az új építésű – jobb épületszerkezeti értékekkel rendelkező – épületek felülreprezentáltak. Ezt a fal, ablak és tető jellemzők esetében lehet megfigyelni.

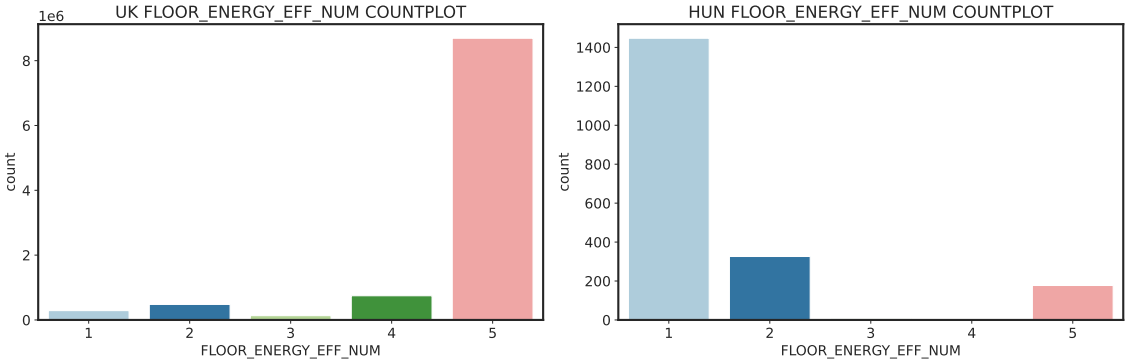


4.5. ábra. Az ablakot leíró jellemző eloszlásának összehasonlítása a két adatbázisban



4.6. ábra. A falat leíró jellemző eloszlásának összehasonlítása a két adatbázisban

Az padlózatot leíró jellemző (4.7) az egyetlen változó, ahol teljesen eltér a két eloszlás. A UK adat itt van a legkevésbé egyenletesen felosztva, szinte az összes érték az 5-ös kategóriába esik. A UK adatbázis dokumentációja szerint felosztott magyar adatok nagy része pedig – pont ellentétes módon – az 1-es kategóriába esik. A padlózatot leíró jellemzőt – az eloszlásbeli különbségek miatt – a későbbi feladatok során nem használtam.

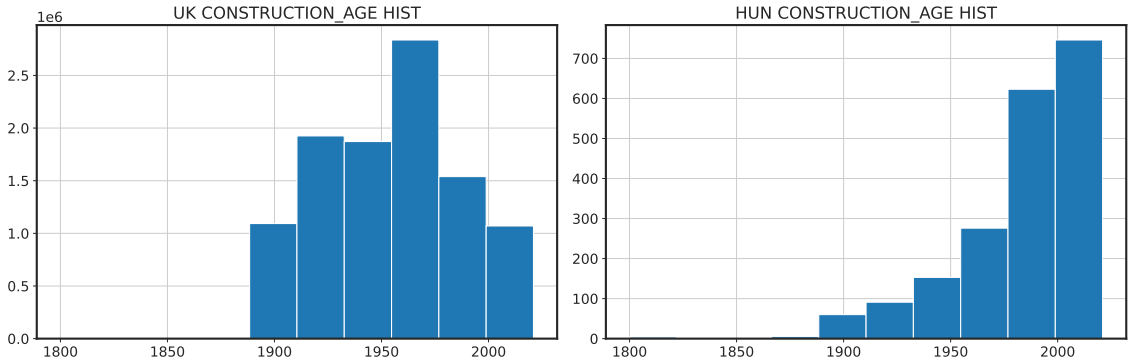


4.7. ábra. A padlózatot leíró jellemző eloszlásának összehasonlítása a két adatbázisban

4.1.2. Folytonos jellemzők eloszlásai

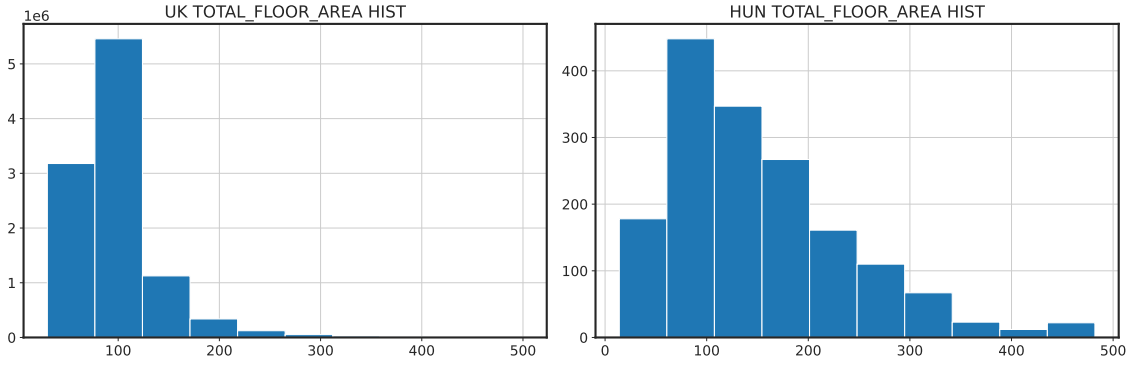
A két adatbázis négy folytonos jellemzőt tartalmaz: építés éve (4.8), teljes alapterület (4.9), belmagasság (4.10) és fűtésfogyasztás (4.11).

Az építés évének eloszlásainál megfigyelhető, hogy a magyar adatbázisban az újépítésű épületek felül vannak reprezentálva. Ez a kvalitatív változóknál észrevett sok 5-ös értéket is igazolja.



4.8. ábra. Az építés évét leíró jellemző eloszlásának összehasonlítása a két adatbázisban

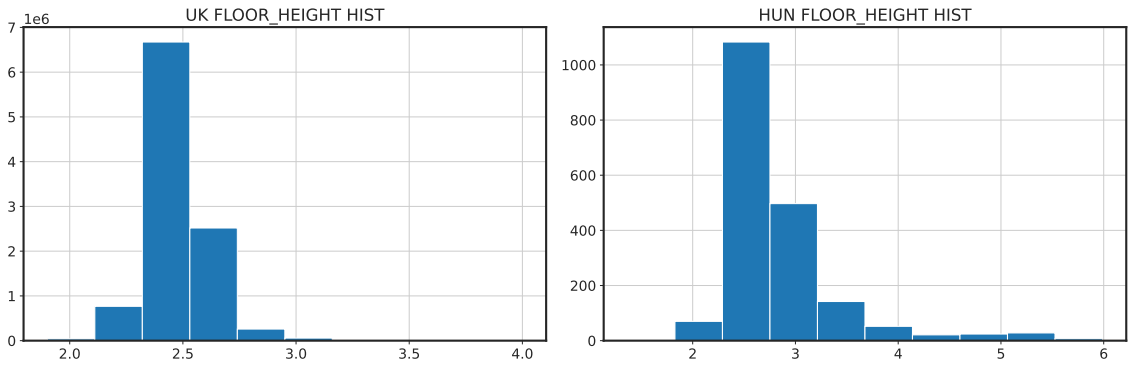
Alapterületet tekintve a UK adatbázis inkább kisebb épületeket tartalmaz, $200m^2$-nél szinte nincsenek nagyobb házak. A magyar épületek eloszlása terebélyesebb, ezt az indokolhatja, hogy a magyar adatbázis tartalmaz intézményi épületeket is, nem csak lakossági ingatlanokat.



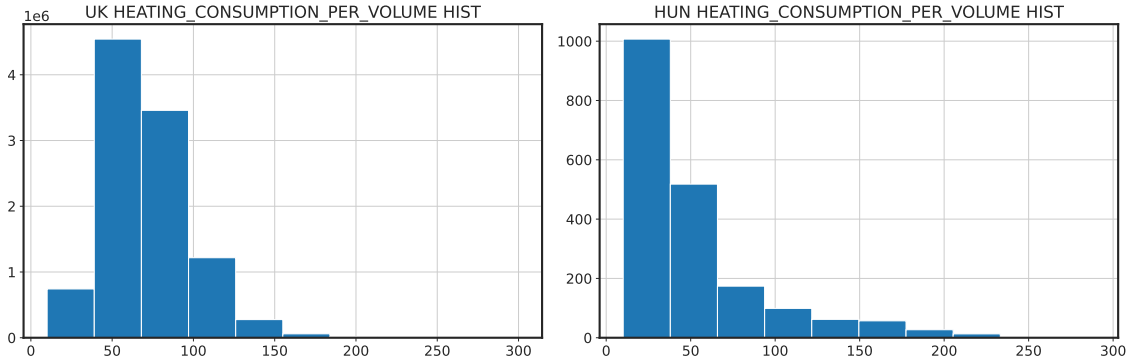
4.9. ábra. Az teljes alapterületet leíró jellemző eloszlásának összehasonlítása a két adatbázisban

A belmagasságokat tekintve nagyon hasonló eloszlásokat láthatunk. A legtöbb belmagasság 2,5m körül van, szintén az intézményi épületeknek köszönhetően a magyar adatbázis tartalmaz néhány kiugró értéket.

A fűtésfogyasztást vizsgálva megfigyelhető, hogy a magyar adatbázis több alacsony fogyasztót tartalmaz. Ez szintén magyarázható azzal, hogy az újépítésű és felújított épületek felülreprezentáltak.



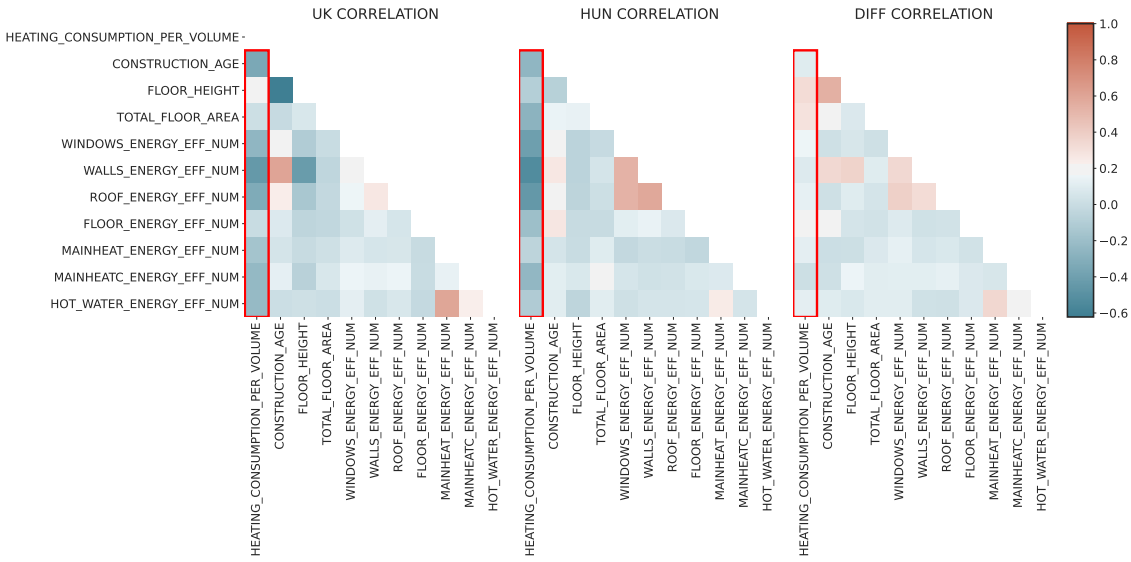
4.10. ábra. Az belmagasságot leíró jellemző eloszlásának összehasonlítása a két adatbázisban



4.11. ábra. Az fűtésfogyasztást leíró jellemző eloszlásának összehasonlítása a két adatbázisban

4.2. Korrelációk összehasonlítása

A feltáró elemzést a korrelációk vizsgálatával folytattam. A két adatkészlet Kendall-féle korrelációit és a korrelációinak különbségeit a 4.12 ábra mutatja be.



4.12. ábra. A két adatbázis korrelációi és korrelációinak különbségei

Az ábrán pirossal bekereteztem a célváltozó (fűtésfogyasztás) korrelációt. A különbségmátrixon megfigyelhető, hogy a célváltozót tekintve nincsenek nagy eltérések. Két kis eltérést figyelhetünk meg a bekeretezett részben, az alapterület és belmagasság nagyobb korrelációt mutat a célváltozóval a magyar adatkészletben. Ez azzal magyarázható, hogy a magyar adatbázis intézményi épületeket is tartalmaz, így e két változó mentén nagyobb szórás figyelhető meg.

A többi változót tekintve a legerősebb eltérést az építés éve és belmagasság tekintetében figyelhetjük meg, amely az építkezési szokásokból eredhet.

Egy nagyobb eltérés-csoportosulás látható még a fal, ablak, tető tekintetében. Ez a hármas a magyar adatkészletben jobban összefügg egymással. Ezt az eloszlások vizsgálata során észrevett különbség magyarázhatja: a UK adatkészlet ezen változói rosszul lettek kvantálva (pl. nincs 5-ös érték).

Felfedezhető még korrelációs eltérés a fűtőrendszer és melegvíz-előállító rendszert jellemző változó között is. A UK adatkészletben jobban összefügg ez a két változó. Ez azzal magyarázható, hogy a több az olyan UK épület, ahol a melegvíz-előállításáért ugyanaz a rendszer felelős, mint a fűtésért. A magyar épületek között több lehet azon rendszer, ahol elektromosan van a melegvíz-előállítva (pl. bojler segítségével).

Az utolsó eltérés a falak energetikai hatékonyság korreláció-különbsége az építés évével, valamint a belmagassággal. Az építés évével való különbséget a felújított épületek felülreprezentáltsága okozhatja (mivel a felújítás éve nem írja felül az építés évét). A belmagasság pedig azért korrelál erősebben a UK adatbázisban a fallal, mivel a fal erősen összefügg az építés évével, ezért egy tranzitív korreláció alakul ki.

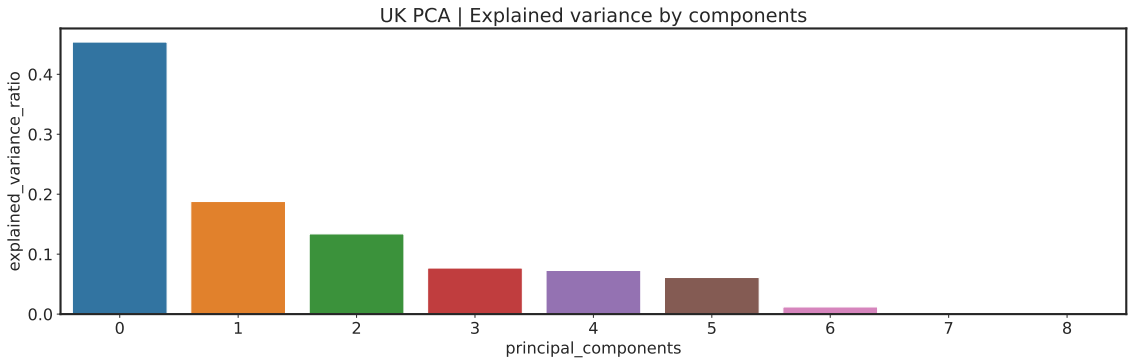
4.3. Főkomponens-analízis

A feltáró elemzést főkomponens-analízissel (Principal Component Analysis, PCA) folytattam. A PCA egy dimenzióredukciós eljárás egy olyan adathalmazon, ahol a változók összefüggenek. Az eljárás során az a cél, hogy úgy állítsunk elő minél kevesebb főkomponenst, hogy a meglévő variancia minél jobban megmaradjon [10].

A főkomponens-analízist kilenc változón végeztem. A padlózat energetikai hatékonyságát leíró változót elhagytam, valamint a célváltozót (fűtésfogyasztást) se vettem bele az elemzésbe.

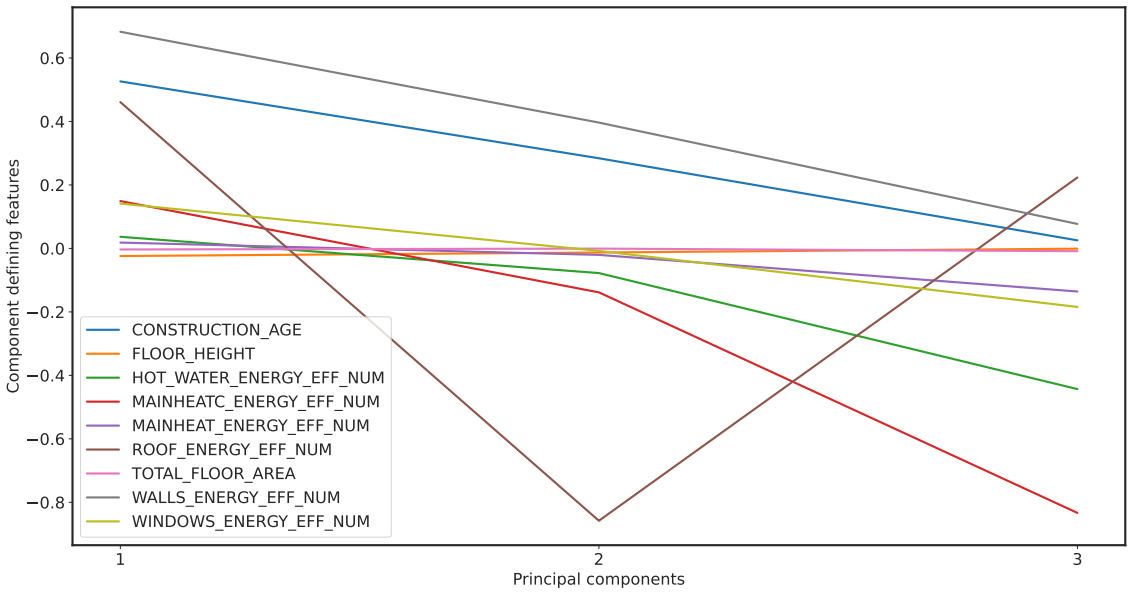
4.3.1. Uk adatbázis főkomponens-analízise

A UK adatbázis főkomponensei szerint leírt varianciákat a 4.13 ábra mutatja be.



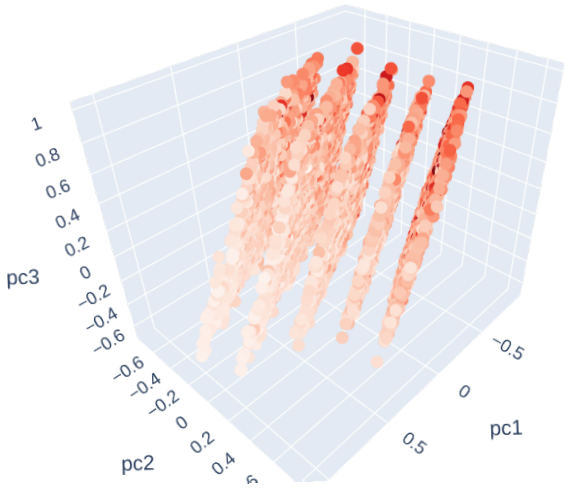
4.13. ábra. UK adatbázis főkomponensei szerint leírt variancia

Az 4.14 ábra mutatja be, hogy az első három főkomponenst melyik változó, milyen mértékben határozza meg. Látható, hogy az első és második főkomponenst az építés éve, a falak és a tető minősége határozza meg. A harmadik komponensnél a fűtésvezérlés és melegvíz előállítás a legdominánsabb változó.



4.14. ábra. UK adatbázis főkomponenseinek összetevői

A három főkomponens térbeli ábrázolását a 4.15 ábra mutatja be, ahol az egyes adatpontok a fogyasztás szerint vannak kiszínezve (minél pirosabb, annál nagyobb). Látható, hogy a fogyasztás egyenletesen növekszik egy térbeli egyenes mentén. Ezenkívül megfigyelhető egy sávos elrendeződés a második főkomponens mentén a kvalitatív változók miatt.

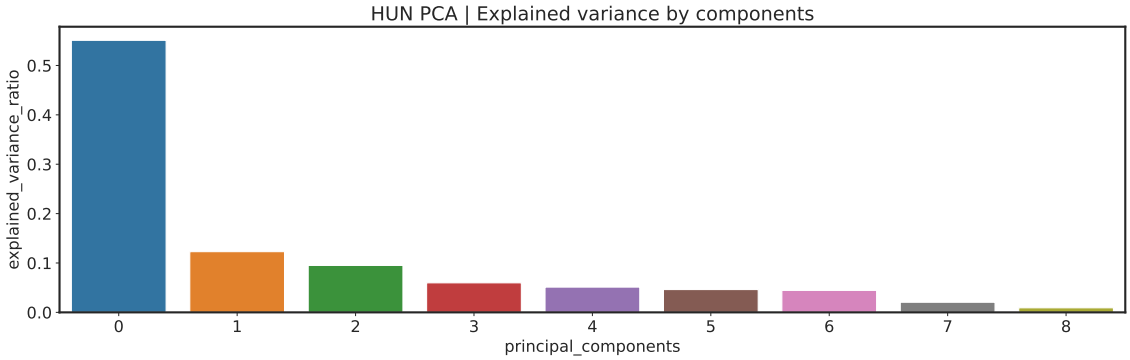


4.15. ábra. UK adatbázis első három főkomponense térben ábrázolva

4.3.2. Magyar adatbázis főkomponens-analízise

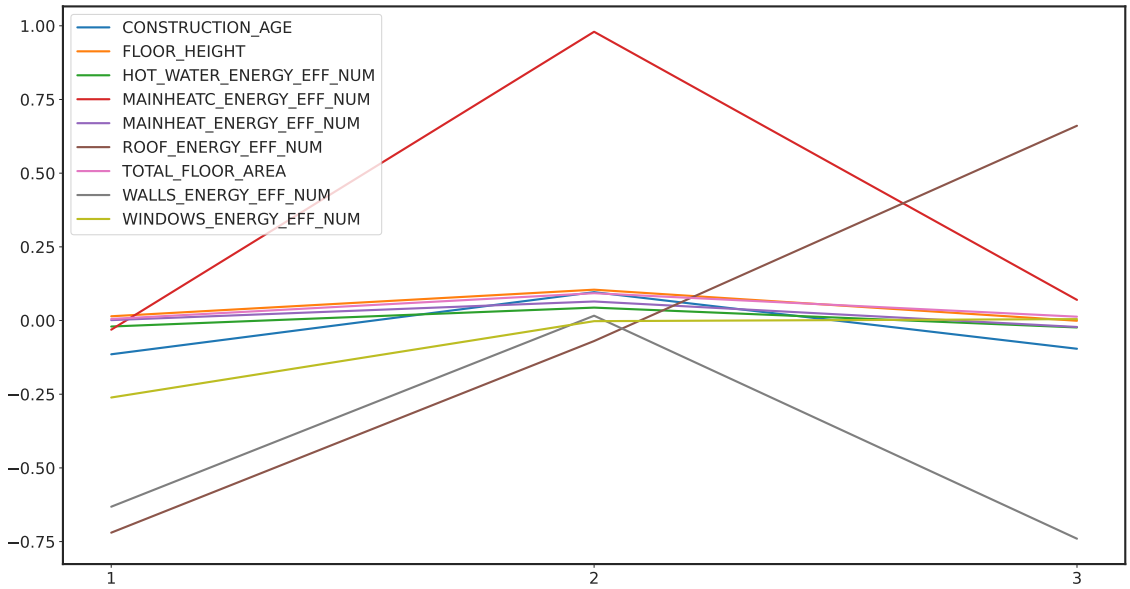
A magyar adatbázis főkomponensei szerint leírt varianciákat a 4.16 ábra mutatja be. A magyar adatkészletben az első főkomponens még erősebb varianciával rendelkezik (eloszlásoknál láthattuk, hogy jobban eloszlanak az értékek), mint a UK-ben látott. Az első komponens ∼55%, míg a második, harmadik összevéve ∼20% varianciát ír le. A három fő-

komponens által leírt variancia nagyjából a UK-beli főkomponensek által leírt varianciának felel meg (∼77%).



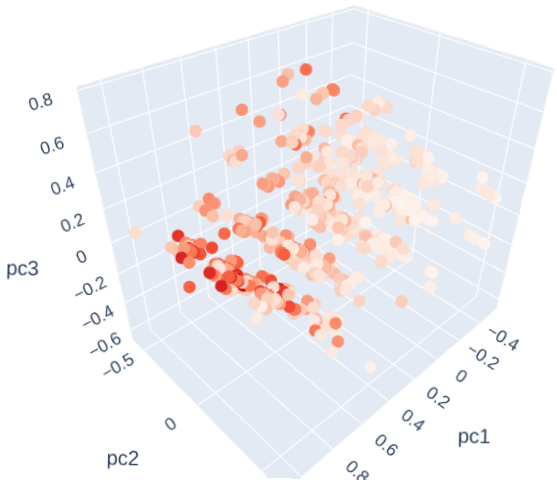
4.16. ábra. Magyar adatbázis főkomponensei szerint leírt variancia

A főkomponensek összetevőit tekintve az első fő különbség a UK-el szemben, hogy az építés éve nem jelentkezik erősen az első főkomponensnél. Ez magyarázható azzal, hogy a 4.8 ábrán látható módon sokkal kisebb a magyar adatokban az építés évének varianciája (sok az újépítésű). A fal és tető fontossága viszont itt is előjön az első komponensnél, mint a UK estében. A másik különbség az, hogy a magyar főkomponenseknél a második és harmadik fel van cserélve a UK-hez képest. A második komponenst írja le a fűtésvezérlés (melegvíz-előállítás nélkül), a harmadik a fal és a tető. Ez pont ellentétes a UK-hez képest.



4.17. ábra. UK adatbázis főkomponenseinek összetevői

Az első három főkomponenst a magyar adatok esetében is vizualizáltam térben (4.18). Ebben az esetben jobban elkülönülő csoportok figyelhetőek meg, és szintén látható átmenet a rossz és jó fogyasztók között.



4.18. ábra. Magyar adatbázis első három főkomponense térben ábrázolva

4.4. A feltáró adatelemzés eredményei

Összességében kijelenthető, hogy a két adatkészlet nagyon sok hasonlóságot hordoz magában. Az eloszlások számos egyezést mutatnak. A kategorikus változókat tekintve a tetőt, fűtést, fűtésvezérélést és melegvíz-előállítást leíró jellemzők hasonló eloszlásokkal rendelkeznek a két adatbázist tekintve. A falak és ablakok energiahatékonyságát jellemző változók eloszlásai eltérnek egymástól. A falak jellemzője a UK adatbázisban főként a hármas és négyes intervallumba esnek, míg a magyar adatkészletben a négyes, ötös tartományba. Az ablakokat tekintve a fő különbség, hogy a UK alig rendelkezik ötös értékekkel, míg a magyar adatkészlet az egyes kategóriában hiányos. A padlózatot leíró jellemző annyira eltér a két adatkészletben, hogy a további munkám során nem használtam fel, mivel valószínűleg más kvantálás szerint történt a változó felosztása a két adatkészletben.

A folytonos jellemzők esetében megfigyelhető, hogy általánosságban újabb építésűek a magyar adatkészletben található megfigyelések. Az épületek alapterületét nézve amagyar adatkészletben viszonylag sok nagyobb alapterületű épület van a UK-hez képest a nem lakossági épületek miatt. A belmagasságok eloszlásai nagyjából lekövetik egymást a két adatkészletben, de megfigyelhető a nagyobb belmagasság felülreprezentációja a magyar-megfigyeléseket tekintve, szintén a nem lakossági épületek miatt. A fűtésfogyasztásokat vizsgálva alapvetően egy pozitív irányú eltolás látható a UK adatbázis részéről a magyarhoz képest.

A korreláció elemzés megmutatta, hogy a célváltozó korrelációi nem térnek el lényegileg a két adatbázisban. A magyar általánosan nagyobb korrelációkkal rendelkezik a célváltozót tekintve, amit a magyar adatkészlet specifikussága indokol (kevesebb különböző épületcsoportot fed le).

A főkomponens elemzésből pedig kiderült, hogy nagyjából ugyanazon változók felelősek az adathalmazok varianciájának meghatározásban. Látható az is, hogy három változóval mindkét adatkészletet tekintve a variancia nagyjából háromnegyede leírható.

Az feltáró adatelemzésből látható, hogy a két adatkészlet logikai szinten nem tér egymástól, a fellelhető összefüggések lényegében megegyeznek (különös tekintettel a célváltozóra). Az elemzés eredményeképpen kijelenthető, hogy egy transzfer tanulási eljárás vizsgálata megalapozott.

5. fejezet

A kvalitatív transzfer-tanítás motivációja

A kutatásom motivációját, a transzfer-tanítás szükségességét a jelen fejezet mutatja be. Az előző fejezetben kiderült, hogy magyar adatbázis is erős korrelációkat mutat az épületeket leíró jellemzők és a fűtésfogyasztás értéke között.

A kiértékelési metrikák után bemutatom, a magyar adatkészleten tanított modell nem tud kellő pontosságú becslést szolgáltatni, mivel nem rendelkezik elég megfigyeléssel. Az utolsó alfejezet pedig bemutatja, a UK adatkészletből épített modell predikciós erejét, valamint a UK modell magyar adatokon való alkalmazását.

5.1. Kiértékelési metrikák

A jelen problémánál nehéz megadni elsőre, hogy mit tekintünk jó eredménynek. Regressziós problémáknál legtöbbször a MAE (Mean Absolute Error) és RMSE (Root Mean Squared Error) metrikákat használják. Ezzel a két metrikával az lehet a probléma, hogy aki nem ismeri az adott szakterületet, nem érzi elsőre, hogy mennyire jó a becslés.

A munkám során a MAE és RMSE mellett kiértékeltem a MAPE (Mean Average Percentage Error) értéket is, ami már beszédesebb lehet egy laikus számára is. Viszonyítási alapként a kapcsolódó kutatások fejezetben bemutatott regressziós megoldást [2] lehet például venni, ahol a kétrétegű megoldással ∼17% MAPE-t értek el.

Ezenkívül létrehoztam egy negyedik metrikát is, amit *CDF Score*-nak neveztem el. Ezt a következőképpen kapjuk meg. Először vesszük a kumulatív eloszlás függvényét a hibának (100%-os hibáig). Ezután ezt ábrázoljuk úgy, hogy az y tengelyen a százalékos hiba mértéke van, az x tengely pedig megmutatja, hogy hány százaléka esik bele az adatoknak az adott százalékos hibába. Végezetül ebből úgy lesz egy kiértékelési metrika, hogy vesszük a függvény alatti területet.

A kiértékeléshez készítettem egy függvényt, ami a predikció és az eredeti érték alapján mindig kiértékeli az eredményt több szempont alapján. A függvény egy kimenete a 5.1 ábrán látható. Az első diagramon a hiba kumulatív eloszlásfüggvénye látható. A második és harmadik diagram a százalékos hiba eloszlását mutatja be (az első abszolút értékben). A negyedik diagram a nominális hiba eloszlását ábrázolja. Az utolsó diagramon egy boxploton ábrázoltam a százalékos hibát. A diagramok alatt sárga háttérrel a négy számszerű kiértékelési metrika (CDF Score, MAPE, RMSE, MAE) látható.

Munkám során alapvetően az MAPE metrikára optimalizáltam, azonban a többi metrikát is figyelembe vettem az optimalizációk során.

5.2. Fogyasztásbecslés a magyar adatok alapján

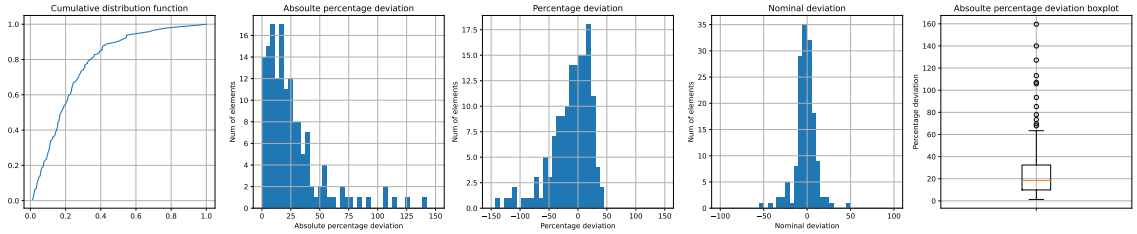
Az épületek fűtésfogyasztásának becslése – jelen esetben – egy regressziós feladat. A becsléskor (6 kategorikus, 3 folytons) bemeneti változót használtam arra, hogy a célvátozóra vonatkozóan (fűtésfogyasztás) minél pontosabb becslést kapjak. A magyar modell építésekor a belmagasságot leíró változót is felhasználtam.

Munkám során több technológiát is kipróbáltam a becslő modell minél pontosabb előállítására. A kipróbált megoldások többek között a következők voltak: XGBoost, CatBoost [21], LightGBM [11], neurális háló - Keras [8]. A számos technológia közül az XGBoost technológia teljesített a legjobban. Az XGBoost egy nyílt forráskódú szoftvercsomag, amely gradiens-boosting keretrendszert biztosít C++, Java, Python, R, Julia, Perl és Scala nyelvi környezetekben.

5.3. A magyar modell kiértékelése

A magyar modellt 2300 megfigyelésen tanítottam be, és 150 – megfelelően kiválasztott – mérésen teszteltem. A tesztmérések kiválasztásánál ügyeltem arra, hogy a tesztadatkészletben ne legyenek a tanítókészelttel túlzottan megegyező mérések. Erre azért volt szükség, mivel a magyar adatkészlet sok olyan megfigyelést tartalmaz, ami majdhogynem megfeleltethető egymásnak (pl. egy épület két alépítményhez készült tanúsítvány).

A betanított modell kiértékelése a 5.1 ábrán látható. A modell átlagos hibája (MAPE) kb. 26,5%, ami a jelen alkalmazást tekintve a tűréshatáron kívül esik. Az elvárt fogyasztás meghatározásánál egy ideális modellt használva – minden mérés esetén – legalább 20%-os hibán belül kell maradni. A hibákat kétféle kategóriára oszthatjuk: a modell az elvárt értéknél nagyobbat, vagy kisebbet becsülhet. Jelen alkalmazás szempontjából az eredeti értéknél nagyobb predikció minőségileg rosszabb. Ennek az az oka, hogy ha egy épülethez nagyobb elvárt fogyasztást rendelünk, mint kéne, akkor előfordulhat, hogy az adott épület nem lesz felülvizsgálva üzemeltetés szempontjából. Ezzel szemben, ha kisebb elvárt értéket határozunk meg, akkor legrosszabb esetben egy felesleges felülvizsgálat történik, viszont nem marad észrevétlen egy esetleges üzemeltetés anomália. Végső soron a túl szigorú elvárt értékeket is elminálni kell, hiszen plusz erőforrás lekötést eredményeznek. A 5.1 ábra középső diagramján látszik, hogy a magyar modell több negatív előjelű hibával rendelkezik, tehát több minőségileg rosszabb hibát produkál.



5.1. ábra. A magyar fogyasztásbecslő modell kiértékelése

Ez azt jelenti, hogy az átlagos hibának még kisebbnek kell lennie. Egy elfogadható megoldást eredményez az is, ha a megfigyelések bizonyos hányadára nem tud a modell becslést adni, azonban hibahatáron belül működik azokra a megfigyelésekre, amelyekre predikciót szolgáltat.

Hiba-intervallum	**0.0 - 0.15**	**0.15 - 0.2**	**0.2 - 0.266**	**0.266 - 0.3**	**0.3 - 0.5**	**0.5 - 0.7**	**0.7 - 0.9**	**0.9 - inf.**
Hun model	58	20	21	7	23	10	3	7

5.1. táblázat. A magyar fogyasztásbecslő modell hibaintervallumai

A 7.4 táblázat a modell által becsült eredmények (balról nyitott) hibaintervallumait mutatja. A táblázatot nézve látszik, hogy a tesztadatok nagyjából egyharmadára kritikus hibát (>=30%) produkál, és pusztán az megfigyelések egyharmadára képes megfelelő becslést adni. A predikciós modell ilyen szempontból kétféle irányba fejleszthető. Egyrészt növelhetjük a modell általános predikciós pontosságát (csökkentve a MAPE értéket), másrészt a kiugró értékek detektálással csökkenthetjük a kritikus és a tűréshatáron kívül eső hibák számát. Jelen dolgozat a második megoldást célozza meg, a későbbiekben bemutatott megoldás segítségével a UK adatok alapján kiugró érték detektálást végzek.

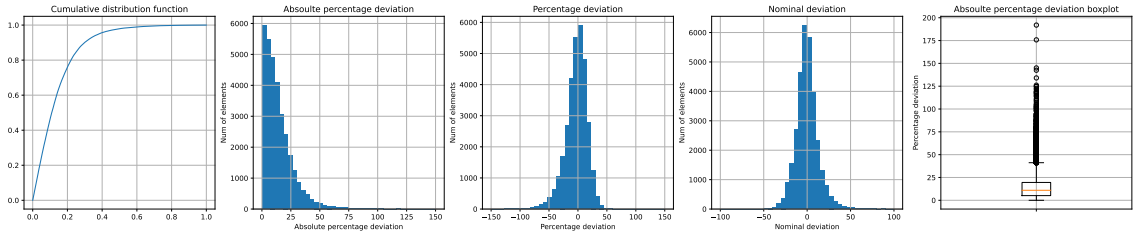
5.4. Fogyasztásbecslés a UK adatok alapján

A UK adatkészlet sokkal terjedelmesebb a magyar adatkészletnél, ezért joggal feltételezhető, hogy a UK adatok alapján tanított modell jobb predikciós eredményt képes produkálni a magyar modellnél. Elsőként megvizsgáltam, hogy a UK adatokból épített modell hogyan teljesít a UK megfigyelésekből leválasztott tesztadat készleten. Ezt követően kipróbáltam, hogy mindenféle transzfer nélkül alkalmazható-e a UK modell a magyar adatokon. Végezetül megnéztem, hogy egy „klasszikus” folytonos transzfer tanítás hogyan teljesít.

5.4.1. Uk modell alkalmazása a UK megfigyeléskre

A UK adatokból épített modellt kb. 220 ezer adaton tanítottam, és 32 ezer adaton teszteltem (azért ekkora adatkészleten, mert a tanító megfigyelések növelése nem mutatott szignifikáns javulást a modell predikciós teljesítményében).

A betanított modell láthatóan sokkal jobban teljesít a UK adatokon, mint az előzőekben bemutatott magyar modell. A 5.2 ábrán látható, hogy 14% körüli a MAPE érték a tesztadatokon. A CDF körbe is szép ívet ír le, és leolvasható róla, hogy a tesztadatok csaknem 80 százaléka 20% hibán belül van.



5.2. ábra. A UK fogyasztásbecslő modell kiértékelése (UK adatokra)

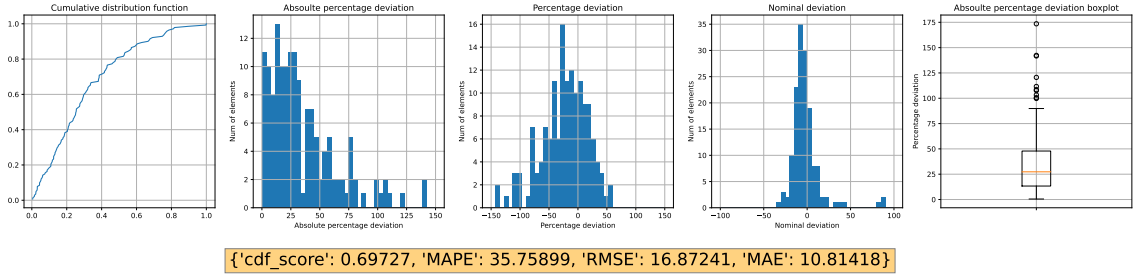
A 5.2 táblázatban látható, hogy a az adatok csaknem kétharmadára megfelelő preidkciót ad a modell, és a nagyjából a becslések 10%-a rendelkezik kritikus hibával. Összességében a UK adatokból felépített modell egy megfelelő alap, ha UK adatokra szeretnénk fogyasztásértéket becsülni.

Hiba-intervallum	0.0 - 0.15	0.15 - 0.2	0.2 - 0.266	0.266 - 0.3	0.3 - 0.5	0.5 - 0.7	0.7 - 0.9	0.9 - inf.
UK model	20406	3920	3431	1086	2508	498	137	52

5.2. táblázat. A UK fogyasztásbecslő modell hibaintervallumai (UK adatokra)

5.4.2. Uk modell alkalmazása a magyar megfigyeléskre

A jelen alfejezetben kipróbáltam, hogy mindenféle transzfer-tanítás nélkül hogyan teljesít a UK modell a magyar adatokon. Mivel a magyar adtok kvalitatív bemeneti változói a UK referenciarendszer szerint vannak felosztva, a folytonos változók (alapterület, belmagasság, építés éve) pedig teljes mértékben összeegyeztethetőek, ezért a model inputálása minden további nélkül megtörténhet. Ebben a predikció fokozott hibáját az okozza, hogy más az input-output hozzárendelés a két adabázisban (más szabályrendszerrel számolnak a UK tanúsítványok előállítása során). Ezenkívül a magyar és UK épülettípusok eloszlása sem megegyező a két adatkészletben.



5.3. ábra. A UK fogyasztásbecslő modell kiértékelése (magyar adatokra)

A 5.3 ábrán látható, hogy kb. 36% MAPE érteket ért el a modell az 149 magyar tesztadaton, ami a magyar modellhez képest elég rossz teljesítmény. A második diagramon és a 5.3 táblázatban viszont látszik, hogy relatív sok jól becsült megfigyelés is van. Az adatok 40 százalékára megfelelő predikciót adott, és aza datok kb. egyharmada 20% hibán belül van. Mindenféle tanítás nélkül ez jó eredménynek mondható, és bizonyítja a két probléma közötti átjárás lehetőségét.

Hiba-intervallum	0.0 - 0.15	0.15 - 0.2	0.2 - 0.266	0.266 - 0.3	0.3 - 0.5	0.5 - 0.7	0.7 - 0.9	0.9 - inf.
UK model (Hun data)	42	12	20	10	30	16	9	10

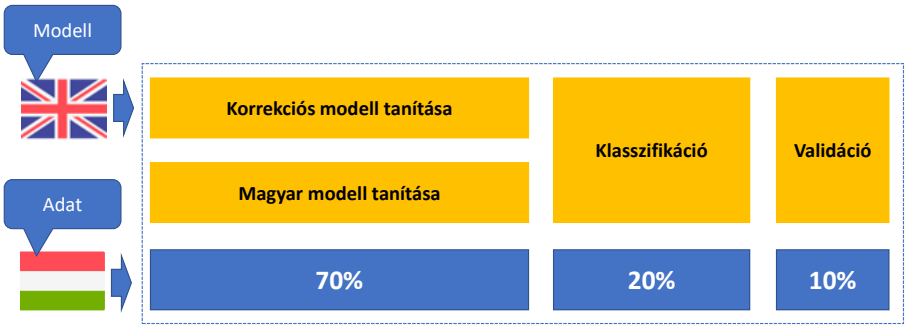
5.3. táblázat. A UK fogyasztásbecslő modell hibaintervallumai (magyar adatokra)

5.4.3. Folytonos transzfer-tanítás megvalósítása

Miután bebizonyosodott, hogy a két adatkészlet között fennáll a transzfer-tanítás lehetősége, egy folytonos transzfer tanítási modellt próbáltam ki. A transzfer tanulással támogatott becslő folyamatot az 7.3 ábra mutatja be.

A magyar adatokat három részre osztottam. Az magyar adatok 70%-án tanítottam be a magyar modellt. Ugyanezen a 70%-on betanítottam egy korrekciós modellt. A korrekciós modell azt hivatott prediktálni, hogy a 5.4 fejezetben bemutatott UK modell tippjéhez mekkora korrekciót kell hozzáadni, hogy az eredeti magyar fogyasztásértéket megkapjuk. A korrekciós modell problámája, hogy elég nagy hibával dolgozik. Azonban a korrekció kiértékelésénél figyelembe kell venni, hogy akár egy nagyon eltérő korrekció is képes javítani

(amennyiben a korrekció előjele megfelelő). Például vegyük azt az esetet, amikor a UK modell 100 kWh fogyasztást tippel, az eredeti érték 200 kWh, a korrekció becsült értéke pedig + 20 kWh. Ebben az estben a hiba 400%-os, mégis javít ahhoz képest, mintha korrekció nélkül használnánk.



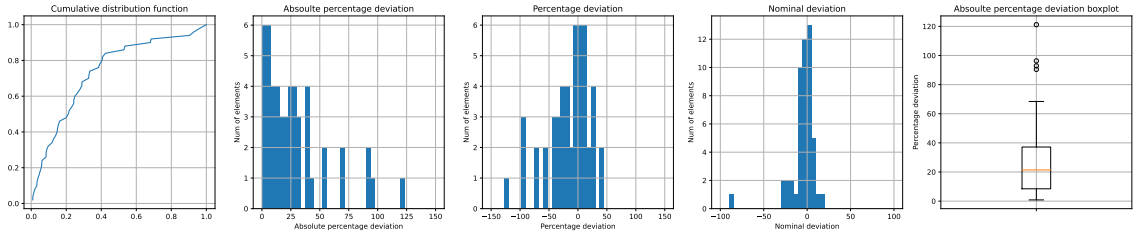
5.4. ábra. Transzfer tanulással támogatott becslő folyamat

A megoldás a következőképpen működik: ha érkezik egy ismeretlen adat, a magyar modell prediktál egy értéket. A UK modell szintén becsül egy értéket, majd a korrekciós modell is visszaad egy előjeles értéket, amit hozzáadunk a UK modell becsléséhez.

A feladat innentől kezdve az, hogy egy ismeretlen adat érkezésekor el tudjuk dönteni, hogy melyik modellt érdemes alkalmazni (magyar modell vagy korrekcióval ellátott UK modell). Ez egy klasszifikációs feladat, amelynek betanítására a magyar adatok 20%-át használtam. Sok esetben csupán kis eltérés van a két modell eredménye között, így azokat az eseteket cimkézetem meg, ahol a UK modell 25%-al jobban teljesít. Így nagyjából az esetek 20%-át cimkéztem fel azzal, hogy a korrigált UK modellt érdemesebb használni.

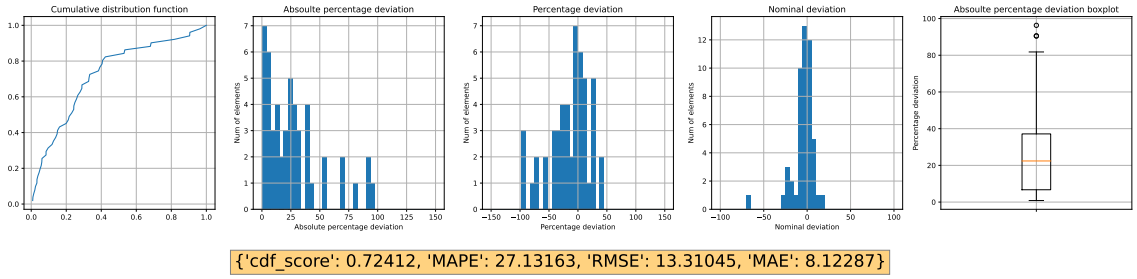
5.4.3.1. Folytonos transzfer-tanítás kiértékelése

Az előző fejezetben bemutatott megközelítést a validációs adathalmazon teszteltem le (a teszadatkészlet a korrekció modell előállításra használtam, ezért választottam le validációs adathalmazt). A 5.5 ábrán látható a magyar modell validációs adatkészleten történt kiértékelése, ez szolgáltatja a referenciaértéket a bemutatott megközelítéshez.



5.5. ábra. A magyar modell kiértékelése a validációs adatokon

A 5.6 ábrán látható a folytonos transzfer tanulás által előállított kompozit modell kiértékelése. Látható, hogy MAPE tekintetében igen csekély eltérés van a két megoldás között. A kompozit modell négy esetben végzett cserét, azaz négyesetben ítélte úgy a címkézés, hogy a transzferrel ellátott UK modell alkalmazása az optimális.



5.6. ábra. A kompozit modell kiértékelése a validációs adatokon

A 5.4 táblázatban látható, hogy a négycsere alkalmával az összes hibaintervallum esetszámát nézve két helyen történt változás (a maradék négy helyen nem javított/rontott annyit, hogy hibaintervallumot lépjen). A megoldás egyik negatívuma, hogy az egyik estben negatív transzfert eredményezett, azaz rontott az eredeti predikcióhoz képest. Az egyik mérés átkerült a 20% hiba fölöttiek osztályába, míg a másik a 90% hibánál nagyobb osztályból a 70-90% közötti osztályba került.

Hiba-intervallum	**0.0 - 0.15**	**0.15 - 0.2**	**0.2 - 0.266**	**0.266 - 0.3**	**0.3 - 0.5**	**0.5 - 0.7**	**0.7 - 0.9**	**0.9 - inf.**
Hun model	20	4	7	3	8	4	0	4
Trans. model	20	3	8	3	8	4	1	3

5.4. táblázat. A magyar és kompozit modell hibaintervallumainak összehasonlítása

A folytonos transzfer-tanulással előállított kompozit modellről összességében kijelenthető, hogy érdemi javítást nem tudott eredményezni. Látható, hogy a megközelítés nem teljesen rossz irány, de ilyenformán éles környezetben nem használható. A megközelítés egyik problémája egyrészt az, hogy a korrekciós modell nagy hibával dolgozik. A másik problémás elem, hogy a klasszifikáció se képes mindig jó cimkézést adni.

A folytonos transzfer tesztelése megmutatta, hogy egy lehetséges transzfer bevezetéséhez az adatkészletek mélyebb – kvalitatív – vizsgálatára van szükség.

6. fejezet

Kvalitatív-kvantitatív transzfer-tanítás

Az előző fejezet eredményei rámutattak arra, hogy érdemes megvizsgálni a két adatkészlet tudás transzfer lehetőségeit. Dolgozatomban egy kvalitatív-kvantitatív transzfer-tanítási megoldást javaslok a – magyar adatok alapján – becsült kiugró fogyasztásértékek detektálására.

A (fizikai) világ folytonos jelenségeinek kvantált változókkal és azok közti kombinatorikus szabályokkal leírása, mint diszciplína a 70-es és nyolcvanas években élte virágkorát *qualitative reasoning* (QR, "kvalitatív következtetés") néven [13]. Az eredeti intuíció az volt, hogy az egyes rendszerleíró folytonos változókat megfelelő vágó-értékekkel ("*landmark*") kvantálva és a rendszer működését leíró parciális differenciálegyenleteket a kvantált változók fölötti szabályokká absztrahálva létrehozható egy olyan kvalitatív modell, melyre a következők igazak.

- A kvalitatív modell teljes rendszerosztályok működését kifejezi logikailag; alkalmazása egy adott rendszerre a landmark-ok, mint paraméterek lekötésével történik. (Ez a szabályokat közvetve befolyásolja, pl. az alapján, hogy egyes landmark-párok esetén melyik érték a nagyobb).
- A kvalitatív modell vizsgálata megvalósítható az informatika diszkrét állapottér-kezelési és vizsgálati eszközeivel (pl. diszkrét szimuláció, szimbolikus végrehajtás, modellellenőrzés, automatizált tételbizonyítás).

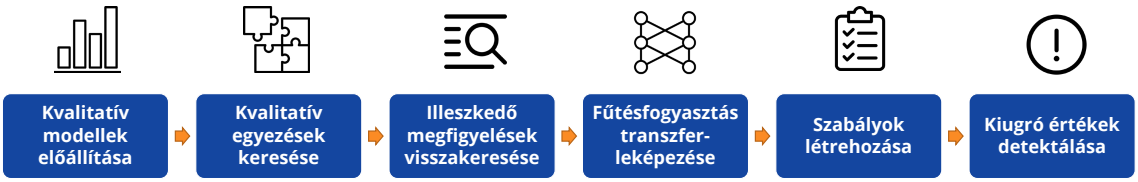
A kvalitatív következkeztetés a fizikai rendszerek kontextusában végül kevéssé terjedt el, köszönhetően a dinamikus rendszerek szimulációinak a kvalitatív absztrakcióból adódó túlzott állapottér-növekedésének köszönhetően. Számos egyéb területen azonban ma is használják (pl. topológiai következtetésekre).

A kvalitatív modellek feletti következtetésre léteznek céleszközök (mint pl. a Garp3 [4]), de ezek alkalmazása legtöbbször akkor indokolt, ha az eredeti QR diszciplínát alkalmazzuk, ahol dinamikus rendszerek leírása történik és a változók első- és második deriváltja is figyelembe vett – kvantált – jellemző. Egyszerűbb, illetve nem fizikai rendszerdinamikai jelenségek elemzését célzó esetekben a kvalitatív "gondolkodást" alkalmazzuk; a QR stílusában kvalitatív modelleket képezve pl. korlátlogikai programozás [23], vagy válaszkészlet-programozás [5] alapú reprezentációhoz.

Szakirodalmi kutatásaim azt mutatták, hogy a kvalitatív modellek alkalmazása a transzfer tanulásban nem egy bevett megközelítés. Az általam vizsgált probléma esetén azonban intuitívan adódik potenciális alkalmazhatósága: várhatóan a különböző energiahatékonysági jellemzők változásának hatása hasonló *jellegű* hatással lesz az energiahatékonyságra különböző országokban, annak ellenére, hogy pl. a kifejezett fogyasztási érték

intervallumok az egyes típusú épületekhez különbözhetnek (pl. különböző energiahasználati, vagy otthontartózkodási szokások miatt). Így potenciálisan követhető az a megközelítés, hogy a ”szabályszerűségeket” egy nagymintás ország alapján (akár részlegesen) kvalitatív modellként megragadva a szabályok egy másik országra alkalmazására mint egy ”landmark-újrameghatározási” probléma tekintünk. A landmark-újrameghatározási probléma egy megoldása közvetlenül alkalmazható a kisebb ország esetén mint kvalitatív értelemben igaznak tekintett szabályrendszer, így az itt bemutatott alkalmazása a magyar modell ellenőrzésére egy *kvalitatív* transzfer tanulási megközelítést ad a transzfer tanulás korábbi definíciója alapján is.

Ebben a munkában egyszerűsítésként feltételezzük, hogy az Egyesült Királyságban és Magyarországon az energiahatékonyságot meghatározó folyamatok kvalitatív értelemben azonosak. Munkám azonban rávilágít arra, hogy maguknak a szabályoknak a feltételes transzfere (azaz transzfer-tanulása) is egy, a jövőben vizsgálatra érdemes lehetőség.



6.1. ábra. A javasolt kvalitatív-kvantitatív transzfer-tanítás kiugró értékek detektálására

Javasolt megoldásom lépéseit a 6.1 ábra mutatja be. A folyamat első lépése a kvalitatív modellek előállítása. Ebben a lépésben mindkét adatbázis folytonos bemeneti változóit, és a célváltozót (fogyasztásérték) is kvantáljuk megfelelő intervallumok mentén. A következő lépésben kiválasztjuk a kvalitatív értelemben megegyező épületcsoportokat, majd visszakeressük a közös épületcsoportokra illeszkedő eredeti megfigyeléseket. Az így megkapott megfigyelések – bizonyos feltételek mentén – összehasonlíthatóvá válnak, és segítségükkel egy transzfer-leképezést tudunk megvalósítani a célváltozóra nézve. A transzfer-leképezést felhasználva a terjedelmes UK adatkészletből egy szabályrendszert tudunk felállítani az egyes épületcsoportok minimum illetve maximum fogyasztását tekintve. Végül az így létrehozott szabályrendszert alkalmazva a kiugró predikciós értékek detektálhatóak lesznek, csökkentve a magyar predikciós modell potenciális hibáit.

6.1. Kvalitatív modellezés és transzfer-tanítás

Energetikai értelemben egy épület jellemezhető egy kvalitatív kategória-kombináció mentén. Például a fal $(Q_1) \to 4$, tető $(Q_2) \to 5$, ablak $(Q_3) \to 4$, fűtésrendszer $(Q_4) \to 3$, stb. hozzárendelés leírja egy épület energetikai hatékonyságát kvalitatív értelemben. A kimenetet szintén jellemezhetjük egy kvalitatív változóval egy megfelelő kvantálás után.

Ebben a kontextusban egy adott épület kvalitatív energetikai viselkedése modellezhető egy relációval, mely minden bemeneti jellemző értékhez a realisztikusnak tekintett kimeneti fogyasztásérték-kvantumok egy halmazát rendeli. Azaz egy adott kategória kombinációhoz tartozhat több fogyasztásérték is ($2^{\mathbb{N}^+}$ egy részhalmaza által elkódolva a kimeneti kategóriák indexeit).

Valósítsa meg egy Q kombinációhoz tartozó output halmaz hozzárendelés egy (φ) leképezés valósítja meg. A több kimeneti érték egyrészt mérési, illetve számítási hibákból, másrészt a nem kezelt egyéb befolyásoló paraméterekből ered. Például a padlózat energiahatékonyságát nem tudtam felvenni a az épületek jellemzésébe, a feltáró adatelemzésben látható okokból. Ugyanígy nem foglalkozom az épületek jellemzése során az épület fizikai formájával, ami szintén fogyasztásbeli különbségeket okoz. Ezek alapján

esetemben a kvalitatív modellek szerkezete a következő:

Kvalitatív modell: $\varphi(Q_1, Q_2, .., Q_k) \rightarrow 2^{\mathbb{N}^+}$

Munkámban részlegesen meghatározott modellek is előű fognak fordulni, de ezek miatt nem szükséges a függvény helyett relációs notáció bevezetése (mondható, hogy amelyik bemenetnél "nem tudunk semmit", ott "minden kimenet lehetséges").

Egy paraméterezett kvalitatív modellről akkor beszélünk, amikor a φ leképezéshez egy O folytonos kimeneti domén felett egy O^L – landmarkok által meghatározott – kvantálás tartozik. A landmarking megvalósít egy particionálást az O output doménre nézve, és megköti, hogy egy adott Q bemenethez tartozó kimenet minimuma legfeljebb akkora lehet, mint a maximuma.

Output domén: $O \subseteq \mathbb{R}$.

Landmarking alapú kvantált domén: $O^L, \forall o_i \in O^L$: $lower(o_i^L) \in \mathbb{R}$ és $upper(o_i^L) \in \mathbb{R}$ és $lower(o_i^L) \leq upper(o_i^L)$. Ezenkívül a kvantálás particionálást valósít meg $min(O)$ és $max(O)$ között.

Paraméterezett kvalitatív modell: $\mathbf{M} = (\varphi, O^L)$

A kutatásom során javasolt és megvalósított kvalitatív modellezési és kvalitatív-kvantitatív transzfer-tanítási megközelítést a 6.2 ábra mutatja be. Elsőként a kvalitatív értelemben vett – kellő előfordulással rendelkező – magyar és UK tartományok közös megfigyelések output eloszlását vizsgáltam. A UK sokkal több illeszkedő megfigyeléssel rendelkezik, így a UK átlagos populáció output eloszlását vettem figyelembe. Mindkét eloszlás tartományon végeztem egy *n-binning* (azonos elemszámú) felosztást, amelyek segítségével előállítottam egy közös – részleges – kvalitatív modellt. Azért részleges, mivel az összes kombinációra nem működik a kvalitatív leképezés, pusztán a közös kombinációkra. A közös kvalitatív modell egy megengedett Q bemenethez hozzárendel egy O^{UK} és O^{HUN} output értéket (kategóriát), az előzőekben véghezvitt *n-binning*-nek megfelelően. A közös kvalitatív modell célja, hogy egy Q bemenethez, megegyező értékű O^{UK} és O^{HUN} értéket rendeljen, ezzel megvalósítva egy transzfer-leképezést a két tartomány között.

Ezt követően az előzőekben létrehozott O^{UK} kategóriákat hozzárendelem az összes UK megfigyeléshez (az előzőekben használt *n-binning*) szerint. A kellő előfordulású kategóriakombinációkat megtartva előállítottam egy közös – szabályrendszert megvalósító – kvalitatív modellt. Az első közös kvalitatív modellhez képest nem egy putput kategoriát rendel a megengdett Q kombinációkhoz, hanem egy kategória intervallumot.

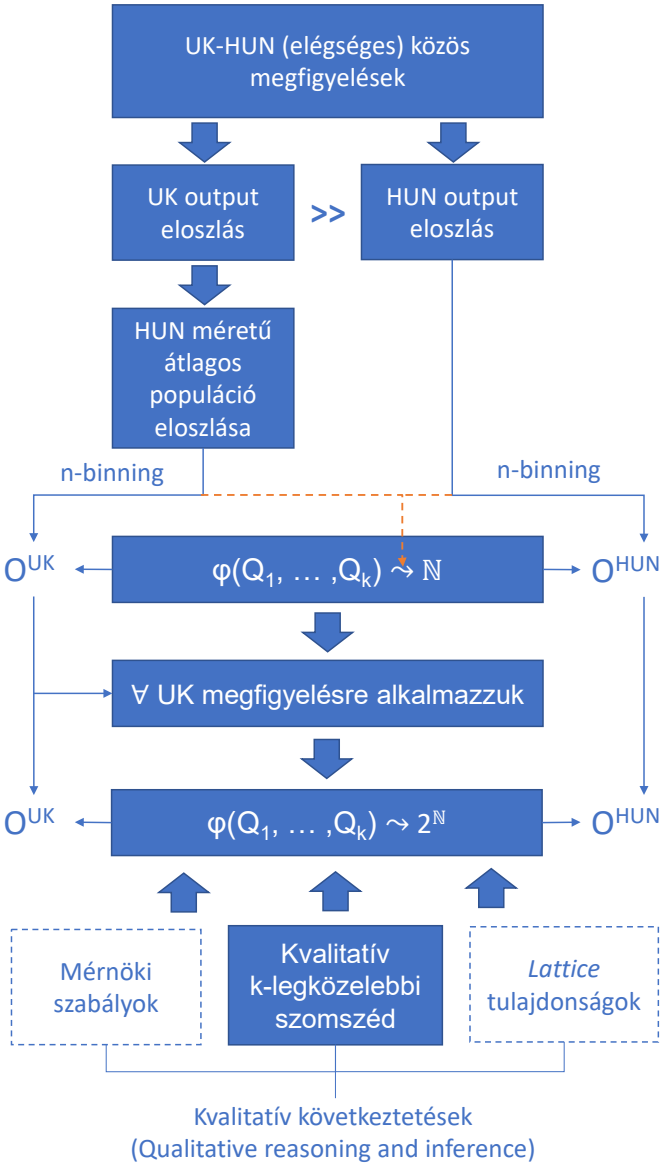
Az így előállított szabályrendszert leíró modell már jóval nagyobb kategória kombináció lefedettséggel rendelkezik, mint csak a közös kategória kombinációkat nézve. Azonban a teljes állapotteret vizsgálva csupán pár százalékos lefedettséget garantál. Ahhoz, hogy a szabályrendszert további input kombinációkra is alkalmazni tudjuk, kvalitatív következtetéseket kell alkalmaznunk. A kvalitatív megközelítés egyik lehetséges fő alkalmazása, hogy pontos kvantitatív értékek nélkül is tudunk következtetéseket megfogalmazni a vizsgált modellen belül. [12]

A dolgozatom további fejezeteiben a k-legközelebbi kvalitatív szomszéd keresést valósítom meg. Amennyiben egy adott Q bemenethez nincs hozzárendelt kimeneti intervallum, akkor a – kvalitatív értelemben – legközelebb eső szabályt alkalmazzuk. A megvalósított modell lehetőséget teremt további kvalitatív következtetési módszerek alkalmazására.

További tervezett kiterjesztése a megközelítésnek, és egyben a kvalitatív modellezés fontos képessége, hogy lehetőség van a modellben mérnöki szabályok közvetlen alkalmazására, illetve kikényszerítésére (a mérnöki szakértői érvelés ismert, hogy jellemzően kvali-

tatív jellegű). Például amennyiben egy szakértő egy adott input kombinációra felső, vagy alsó korláttal tud szolgálni, akkor ez a tudás egyszerűen hozzáadható a modellhez.

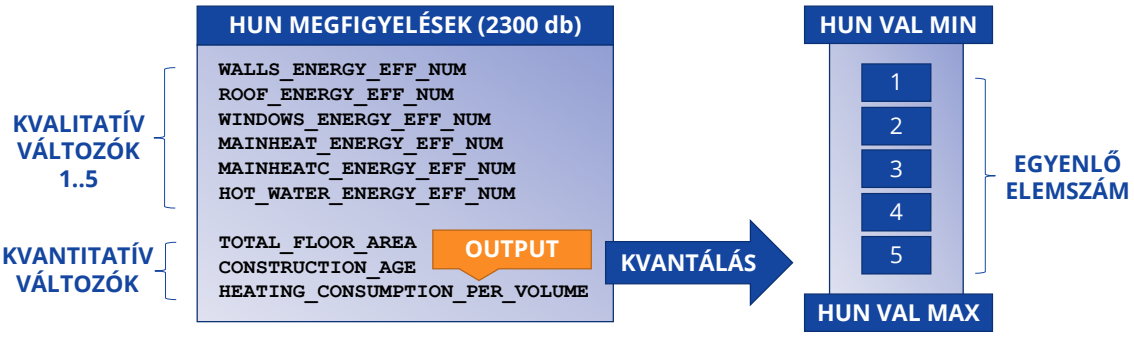
A még a UK adatok alapján is viszonylag alapcsony közvetlen bemeneti kategória-fedésű kvalitatív modell *következtetett* kategória-kimenetekkel "feltöltését" is tervezem. Ez várhatóan lehetséges a bemenetek feletti részleges rendezés reláció hatásának a kimenetek egymáshoz képest elhelyezkedésére való befolyásának megfogalmazásával (pl. "jobb" épület nem fogyaszthat "többet"), az Allen-féle intervallum-logika mentén.



6.2. ábra. Kvalitatív modellezés és transzfer-tanítás

6.2. Kvalitatív modellek előállítása

A modellek előállítása előtt a 2500 magyar megfigyelésből 200 mérést levágtam tesztelés céljából. A 200 adatsorból 150 megfigyelés képezi a teszt adatkészletet, és a maradék 50 megfigyelés validációs célokat szolgál.



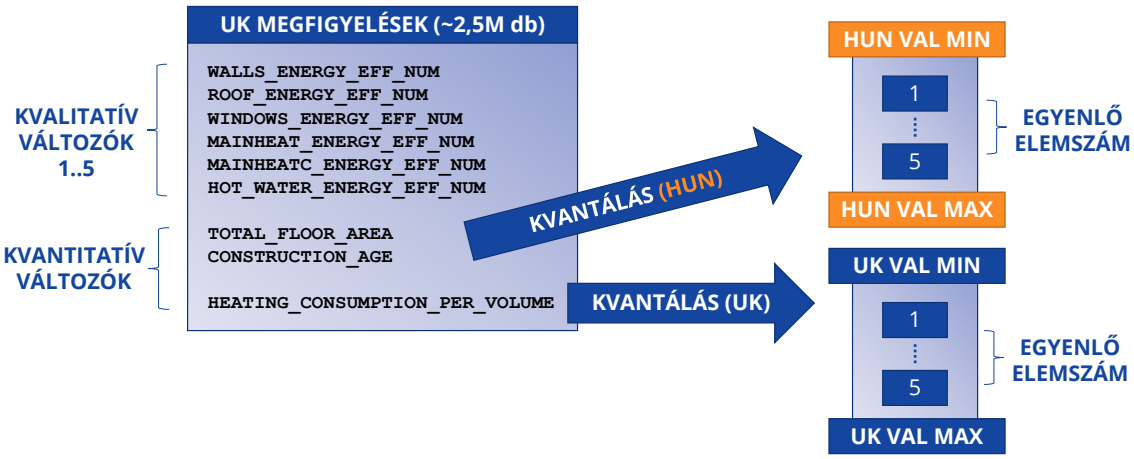
6.3. ábra. A magyar kvalitatív modell előállítása

A magyar kvalitatív modell előállítását a 6.3 ábra mutatja be. A magyar megfigyelésekből álló input modell hat, az előző fejezetekben bemutatott kvalitatív jellemzővel rendelkezik. A maradék három használt kvantitatív változó: alapterült (m^2), építés éve és fogyasztásérték (kWh). A belmagasságot nem vizsgáltam, mert úgy ítéltem meg, hogy az nem játszik olyan nagy szerepet az épületek disztingválása során. A kvalitatív modell előállítása során mind a három kvantitatív változót – az eredeti kvalitatív változókhoz híven – egytől ötös skálán osztottam fel. Az intervallumokat úgy határoztam meg, hogy minden értéhez egyenlő elemszám tartozzon.

	1	2	3	4	5
Alapterület	87,83	138,91	209,54	666,7	41417,25
Építés éve	1964	1974	1988	2002	2021
Fűtésfogyasztás	24,28	32,86	44,059	72,95	289,02

6.1. táblázat. A magyar adatkészlet folytonos változóinak kvalitatív felosztása

A kvantálás során felhasznált intervallumok maximum (jobb oldali) értékeit a 6.1 táblázat mutatja be. Látható, hogy a két szélső kategóriát nézve elég nagy különbségek is előfordulnak. Például az alapterületet tekintve, ha két épületet vizsgálunk, egyet az első, másodikat az utolsó kategóriára nézve, akkor majdnem ötszázszoros különbség is előfordulhat. Ez számunkra azért kedvező, mert így jól disztingválhatóak egyes épületek.



6.4. ábra. A UK kvalitatív modell előállítása

A magyar adatkészlet kvantálása után a UK adatkészletben is előállítottam a kvalitatív változókat. A UK adatbázis kvvantálását a 6.2 ábra mutatja be. Ebben az esetben

kétféle felosztást végeztem. Az input paramétereket (alapterület, építés éve) a 6.1 táblázatban bemutatott intervallumok szerint osztottam fel. Ezzel a felosztással az egyes kvalitatív kombinációk összehasonlíthatóvá válnak a két adatbázist tekinte. Az output paraméteret, vagyis a fogyasztásértéket a UK szerinti eloszlás alapján osztottam fel, minden kategóriába egyenlő elemszámot véve. A kvantálás során felhasznált intervallumok maximum értékei a 6.2 táblázatban láthatóak.

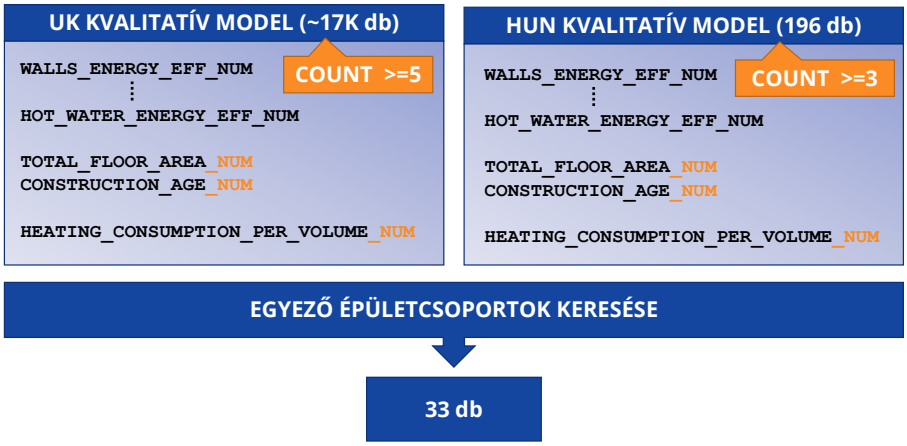
	1	**2**	**3**	**4**	**5**
Fűtésfogyasztás	49,77	60,6	72,56	89,71	499,52

6.2. táblázat. A UK adatkészlet folytonos változóinak kvalitatív felosztása

Megfigyelhető, hogy az egész adatkészletet nézve a UK épületek nagyobb fogyasztás értékekkelrendelkeznek. Persze ez így önmagában nem hordoz információt, mivel az eltérést a jó, illetve rossz épületek eloszlása is befolyásolja. A kimeneti értékek összehasonlításra megegyező kvalitatív kombinációkat kell vizsgálnunk, ezt a vizsgálatot taglalja a következő alfejezet.

6.3. Kvalitatívan megegyező épületcsoportok és épületek

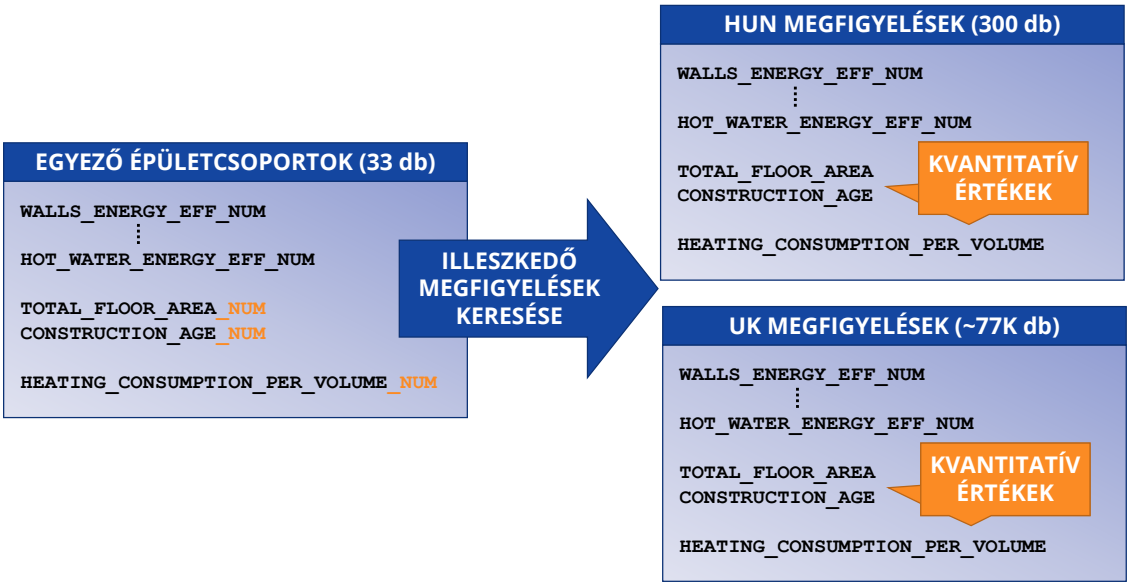
A megoldás következő részében a kvalitatívan megegyező épületcsoportokat kerestem, amit a 6.5 ábra mutat be. Elsőként a két kvalitatív modellből csak azokat a megfigyeléseket tartottam meg, ahol UK oldalon legalább öt, magyar oldalon legalább három előfordulás volt. Ezt azért csináltam, hogy a nagyon ritka épületcsoportokat ne torzítsák a későbbi eredményeket. Így a UK oldalon kb. 17 ezer, míg magyar oldalon 197 épületcsoport keletkezett. Következő lépésként az épületcsoportokat tartalmazó két adatkészletet az input változók mentén összefűztem, és így 36 darab közös épületcsoportot találtam.



6.5. ábra. Kvalitatívan megegyező épületcsoportok keresése

A folyamat következő lépéseként a közös épületcsoportokra illeszkedő megfigyeléseket kerestem meg mindkét adatbázisban, az 6.6 ábra alapján. A 33 darab közös épületcsoportra 300 megfigyelés illeszkedett a magyar adatbázisban, illetve nagyjából 77 ezer a UK adathalmazon. Jól látható, hogy a UK adathalmazra sokkal több illeszkedő mérés esik (a UK adathalmaz terjedelme miatt).

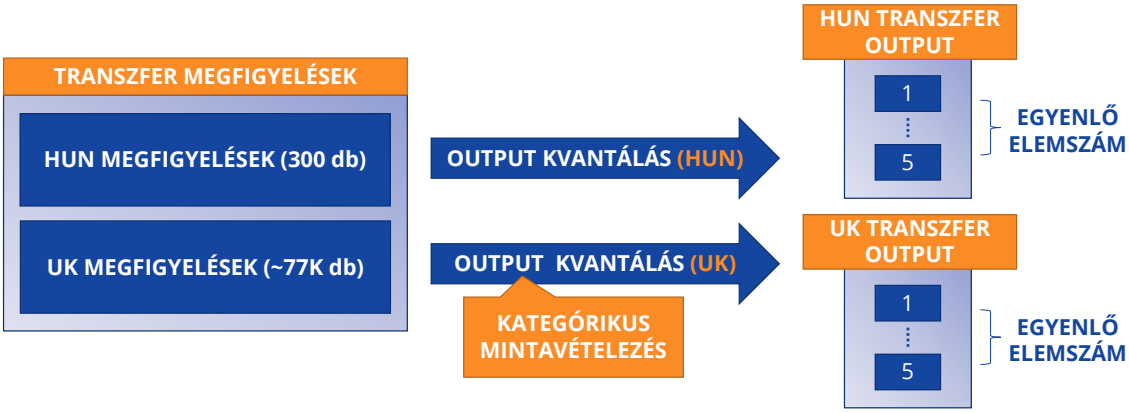
Az összes kategóriakombinációra (33 db) nézve a UK adatkészlet legalább kétszer annyi mérést tartalmaz, de van olyan kategória is, ahol ezerszer annyi adatot tartalmaz (a UK) egy adott kvantitatív kombinációra nézve (HU-hoz képest).



6.6. ábra. Illeszkedő megfigyelések keresése

6.4. Fűtésfogyasztás transzfer-leképezése

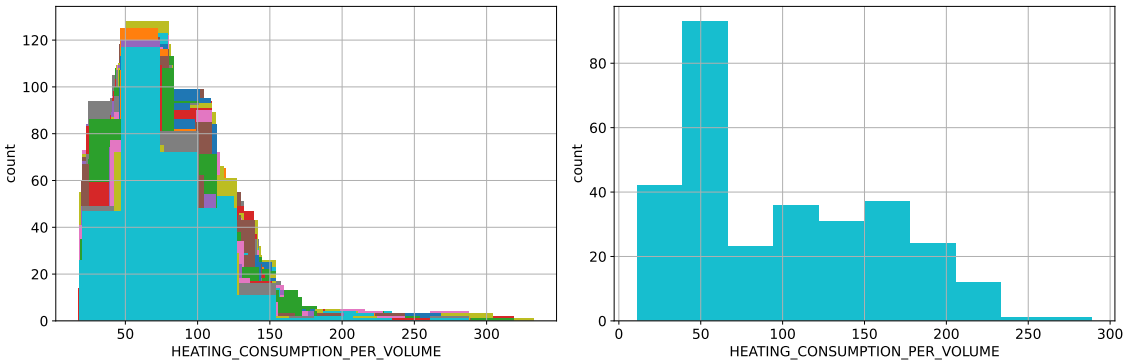
A folyamat következő lépésében történik a tényleges transzfer-tanítás megvalósítása, amelyet a 6.7 ábra mutat be. A megoldás alapgondolata, hogy egyező kategóriakombinációk segítségével vizsgáljuk a megtalált (transzfer) megfigyeléseket, és a saját output eloszlásuk alapján felosztjuk őket öt (egyenlő elemszámú) kategóriára.



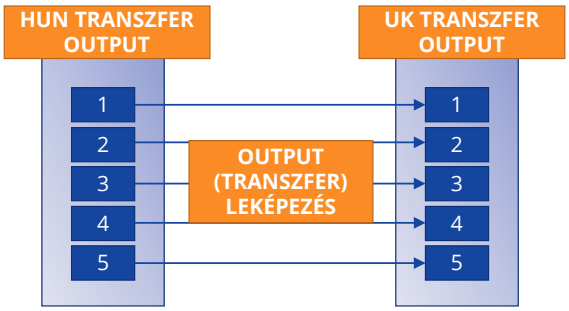
6.7. ábra. Transzfer outputok előállítása

A magyar adatok esetén ez egyszerűen történik, a megtalált 300 megfigyelés alapján adódik az egytől ötös felosztás. A UK adatbázisban sokkal több közös megfigyelést találtunk, ezért ott kategorikus mintavételezést alkalmaztam. A UK adatbázisból a magyar kombinációknak megfelelő elemszámú (összesen 300 darabot) megfigyelést mintavételeztem, majd erre vizsgáltam a kimenet eloszlását. Így már egyenlő elemszámú megfigyelések kimenetét tudtam összehasonlítani. A mintavételezési hiba csökkentése érdekében a mintavételezést ezerszer elvégeztem, és átlagoltam az intervallumok határait. A mintavételzés során előállt UK hisztogrammok (bal) és a magyar adatok alapján előállt hisztogram (jobb) a 6.8 ábrán figyelhető meg.

A leképezés ezután egyszerűen történik a 6.9 ábra alapján. A leképezés során a megegyező kimeneti értékeket rendeltem össze (1→1, 2→2, stb.). Ez lényegében azt jelenti,



6.8. ábra. Fűtésfogyasztások eloszlásai (bal: UK, jobb: HUN)



6.9. ábra. Output transzfer-leképezés

hogy mely kimeneti értékeket tekintünk rossznak, és jónak a magyar, illetve a UK viszonyítási rendszert tekintve. Például, ahogy a 6.3 táblázatban is látható, a magyar rendszerben a maximum 45,1 kWh/m3/év, míg a UK rendszerben ez 49,48 kWh/m3/év. Látható, hogy az első két kategória intervallumaiban nincs túl nagy eltérés, vagyis a jó fogyasztók hasonló besorolást kapnak. A harmadik kategóriától nézve viszont nagyobb eltérések fedezhetőek fel a kvantálás során.

	1	**2**	**3**	**4**	**5**
HUN transzfer output	45,1	56,67	109,64	164,55	289,02
UK transzfer output	49,48	63,55	81,35	101,57	251,47

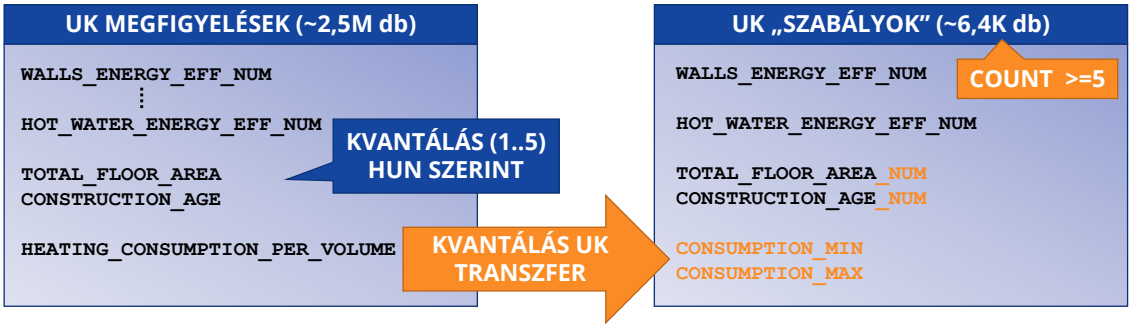
6.3. táblázat. Transzfer output értékek összehasonlítása

Az eredeti – teljes adathalmazon felosztott – intervallumokat nézve a UK intervallumok (6.2) nem térnek el nagyságrendileg. A magyar intervallumokat nézve (6.1) viszont szinte az összes intervallum jobb oldali értéke a duplájára nőtt.

6.5. Kiugró értékek detektálása

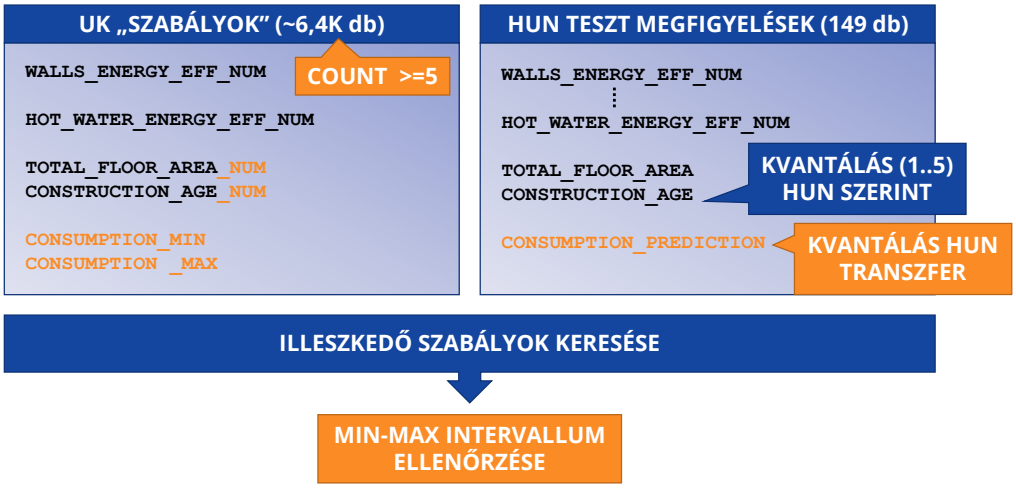
A transzfer leképezés előállítása után egy predikció felülvizsgálatra alkalmas szabályrendszert állítottam fel a 6.10 ábra szerint.A szabályrendszer hozzárendel egy adott kategóriakombinációhoz egy minimum, illetve maximum fogyasztás kategória értéket (A 6.1 alfejezetben említettem, hogy egy kombinációhoz több fogyasztásérték is tartozhat).

A szabályrendszert úgy állítom elő, hogy a UK méréseket a UK transzfer output (6.3 táblázat) szerint kvantálom (az inputokat HUN szerint), és a legalább öt előfordulással rendelkező méréseket megtartva aggregálom a kategóriakombinációkat minimumra és maximumra nézve. A létrejött szabályok közöl eldobtam még azokat, amelyek minimum



6.10. ábra. A szabályrendszer előállítása

értéke egy, és a maximum értéke öt, mivel ezek nem hordoznak plusz információt a kiugró értékek keresését tekintve. Az aggregálás során nagyjából 6400 szabályt hoztam létre, ami azt jelenti, hogy 6400 kategóriakombinációra (épületcsoportra) képes a rendszer egy megengedett output tartományt meghatározni.



6.11. ábra. Kiugró értékek detektálása

Az előállított szabályrendszer segítségével már detektálhatóak a kiugró predikciós értékek a 6.11 ábra szerint. A tesztelésre használt megfigyelések becsült fogyasztását a HUN transzfer output (6.3) szerint kell kvantálni (az inputokat HUN szerint). A tesztadatok kvantálása után a szabályok illesztése következik, ahol megvizsgáljuk, hogy egy adott kategóriakombinációra nézett fogyasztás-predikció a minimum és maximum intervallumokba esik-e.

6.6. A kvalitatív szabályrendszer kiterjesztése

Az előzőekben bemutatott szabályrendszerek egyik hiányossága az lehet, hogy nem képesek kellő nagyságú kategóriakombinációt lefedni. Az előző alfejezetben láthattuk, hogy egy relatív nagy méretű (kb. 2,5 millió) adatbázis esetén is viszonylag kis állapotteret tudunk lefedni a generált szabályokkal. A teljes állapottér 5^8 (390 625) kombinációból áll (mind a nyolc kvalitatív bemenethez tartozik öt lehetséges érték). A fent előállított szabályrendszer nagyjából 6400 szabályból áll, ami a teljes állapottér kevesebb, mint 2%-a. Ezek szerint az esetek 98%-ában nem tudunk szabályt illeszteni a megfigyelésre. Egy másik probléma

az lehet, hogy habár sikerül szabályt illesztenünk, az adott szabály túl megengedő (pl. 1-4 intervallum), ezért nem szűri ki a potenciális kiugró értéket.

A felvázolt problémák elkerülése végett egy lehetséges megoldás a meglévő kvalitatív szabályrendszer kiterjesztése a le nem fedett megfigyelésekre is. A javasolt megoldásban a szabályrendszer kiterjesztése a kvalitatív értelembe vett leghasonlóbb illeszkedő szabály megkeresésével történik. A gondolatmenet alapelve, hogy ha létezik egy szabály, amely csupán egy értékben, és egy változó mentén tér el egy megfigyeléstől, akkor feltételezhetően az adott szabály érvényes lehet az adott megfigyelésre. Például ha veszünk két belvárosi lakást ugyanazon energetikai paraméterekkel, és csupán az alapterületben van eltérés – pl. 60 és 80m^2 –, akkor az adott szabály elméleti síkon alkalmazható.

A hasonló szabályok megtalálására kényszerprogramozást (constraint programming, CP) alkalmaztam. A CP használatával – a hasonlóságon kívül – egyéb kényszerek is meghatározhatóak (például a szabályok intervallum hossza). A kényszerprogramozást a következő alfejezet mutatja be röviden.

6.6.1. Kényszerprogramozás

Kényszerprogramozás vagy korlátkielégítési probléma (Constraint Satisfaction Problem, CSP), során lehetséges megoldásokat keresünk egy – általában – nagyon terjedelmes megoldás-jelölt halmazból, ahol a probléma definiált kényszerek által modellezhető. A kényszerprogramozást jelenleg számos területen alkalmazzák, többek között: ütemezés tervezés, útválasztási problémák, hálózati problémák és bioinformatikai feladatok. [23]

Egy kényszerprogramozással felírt problémát akkor tekintünk megoldottnak, ha probléma során definált összes változóhoz sikerül értéket rendelni úgy, hogy a hozzárendelések minden kényszernek eleget tesznek. A CP kereső algoritmusok alapelve, hogy keresési térben nagy vágásokat képes végezni azáltal, hogy felismeri azokat a változó-érték kombinációkat, amelyek sértik a kényszereket. A CP ezáltal alkalmas nagy méretű, komplex problémák hatékony megoldására.

6.6.2. A kényszerprogramozás definíciója

Egy kényszerprogramozási probléma három komponensből áll: $\mathbb{X}$, $\mathbb{D}$ és $\mathbb{C}$. Az $\mathbb{X}$ egy n nagyságú változó halmaz: $\mathbb{X} = \{\mathbb{x}_1, .., \mathbb{x}_n\}$. A $\mathbb{D}$ a változókhoz tartozó tartomány halmaz: $\mathbb{D} = \{\mathbb{D}_1, .., \mathbb{D}_n\}$. A $\mathbb{C}$ pedig egy kényszerhalmazat definiál, ami megadja a lehetséges érték hozzárendeléseket.

Minden $\mathbb{D}_i$ tartomány $\{\mathbb{v}_1, .., \mathbb{v}_k\}$ megengedett értékekből áll az $\mathbb{X}_i$ változóra nézve. Minden kényszer egy $< scope, rel >$ párosból áll, ahol a *scope* azon változók halmaza, akik részt vesznek a kényszerben, a *rel* reláció pedig meghatározza, hogy a kényszerben szereplő változók milyen értéket vehetnek fel. A reláció reprezentálható egy explicit változó érték hozzárendelés listával, ahol konkrétan fel vannak sorolva a lehetséges kombinációk, vagy egy absztrakt relációval is.

Egy CP probléma megoldásához definiálnunk kell a megoldás állapotterét is. Minden állapot egy hozzárendelés: $\{\mathbb{X}_i = \mathbb{v}_i, \mathbb{X}_j = \mathbb{v}_j, ..\}$. Ha egyik feltétel sem sérül a hozzárendelések során, akkor konzisztens megoldásról beszélünk. Teljes hozzárendelésről akkor beszélünk, ha minden változóhoz sikerült értéket rendelni, ellenkező esetben részleges hozzárendelésről beszélhetünk. [24]

6.6.3. A szabályrendszer kiterjesztése kényszerprogramozással

A dolgozatban bemutatott megközelítésben egy kényszerprogramozással előállított modellt használtam a kvalitatív szabályrendszer kiterjesztésére. A megoldás során a Google OR-Tools Constraint Programming [19] Python csomgját használtam.

A model minden kvalitatív változó számára egy egészértékű változót hoztam létre (*NewIntVar*), amely egytől ötig vehet fel értéket (6.12, 2. sor). Ezek a változók fogják tárolni a kiválasztott szabály kategorikus jellemzőinek értékeit. Az összes létrehozott változóhoz hozzárendeltem, hogy egy adott indexhez csak a hozzátartozó változóérték rendelhető (*AddAllowedAssignments*) minden kategóriára nézve (6.12, 3. sor). Ezzel kikényszerítve, hogy ha egy szabály attribútumát kiválasztja, akkor az összes többi attribútumát is ki kell választania. Ezt követően minden kategorikus jellemzőhöz készítettem egy egészértékű változót (6.12, 4. sor), amely az adott jellemző és a megfigyelés különbségét tartalmazza (ez egytől négyig vehet fel értéket).

```
1  # Fal jellemzo (szabaly - megfigyeles) ekvivalencia kikenyszeritese
2  walls_val = model.NewIntVar(1, 5, 'walls_val')
3  model.AddAllowedAssignments([index,walls_val], walls_tuple)
4  model.AddAbsEquality(hun_walls, walls_val)
5
6  # Teto jellemzo (szabaly - megfigyeles) kulonbseg valtozo letrehozasa
7  roof_val = model.NewIntVar(1, 5, 'roof_val')
8  model.AddAllowedAssignments([index,roof_val], roof_tuple)
9  roof_diff_abs = model.NewIntVar(0, 4, 'roof_diff_abs')
10 model.AddAbsEquality(roof_diff_abs, hun_roof- roof_val)
11 #...
12 # Minimalizalasi feltetel meghatarozasa
13 # model.Minimize(roof_diff_abs + ...
```

6.12. ábra. Kódrészlet a kényszerprogramozás megvalósításából

A modell optimalizálási feltétele, hogy minimalizálja a különbségeket tartalmazó változók értékeit (6.12, 13. sor). Az optimalizáció során akár súlyozhatjuk is az egyes különbség-változókat (például ha a tető jellemzőben van eltérés, azt nagyobb súllyal vegye bele az optimalizációba). Ez azért lehet előnyös, mert a fogyasztás szempontjából kevésbé disztingváló jellemzők kevésbé fontosak a szabálykeresés során. A felállított modell segítségével megtalálhatjuk mindig a legközelebbi szabályt a megfigyelésekhez, és a súlyozás segítségével ezt még finomra is hangolhatjuk.

A bemutatott kényszerprogramozási modellt kiegészítettem két plusz feltétellel. Az egyik megkötés, hogy a fall jellemzőben nem engedek meg eltérést (6.12, 10. sor), mivel – a korreláció elemzésből is fakadóan – túlságosan befolyásolja a kimenti értéket. A másik feltétel, amit megkötöttem, hogy csak olyan szabályt fogadok el, ahol a predikció a szabály intervallumán kívül esik (*OnlyEnforceIf*). Ez egy nagyon erős feltétel, viszont kikényszeríti azt, hogy kiugróértékeket találjunk. Ez ilyen szabály bevezetése azt jelenti, hogy minden predikcióra ráilleszt egy ellentmondó szabályt. Azonban ez azért nem probléma, mivel a szabály és a mérés közötti "távolság" alapján szabályozhatjuk, hogy milyen "távoli" szabályokat fogadunk még el.

A model visszaadja az összes teszt megfigyelésre a hozzá legközelebb lévő – és az extra feltételeknek megfelelő – szabályt, az egyes jellemzők menti eltéréseket, valamint az összeltérést (optimalizációs értéket). A optimális szabálykeresés teljes kódrészlete megtalálható a függelékben (F.1, F.2).

7. fejezet

A javasolt módszer alkalmazása

A jelen fejezet bemutatja a 5.4.3.1 alfejezetben ismertetett hibrid kvalitatív-kvantitatív transzfer-tanítással előállított kiugró érték detekció eredményét. A javasolt módszer kiértékelést elsőként a tesztadatokon végeztem el. A tesztadatok segítségével finomra hangoltam a kiugró érték keresést, majd a végső megoldást a validációs adatokon teszteltem.

A kiugró érték keresés során bizonyos méréseket eldobunk, amennyiben nem illeszkedik a becsült érték a szabályrendszerre. A kiugró érték keresése kiértékelésekor meg kell vizsgálni, hogy az eldobott becslések hibája mekkora volt. Ha a tűréshatáron kívül esik a mérés hibája, akkor egy jó detektálásról beszélhetünk. Azonban vannak esetek, amikor jó (azaz tűréshatáron belüli hibával rendelkező) predikciót dobunk el. Ilyenkor egy hamis detekcióról beszélünk, amely csökkenti a modell predikciós erejét.

A hamis detekció egy másik fajtája, amikor habár a hiba mérték nagy, azonban a szabály illesztés logikailag nem helytálló. Ez abban az esetben lehet, ha a tényleges fogyasztás kívül esik a szabályrendszer által megadott intervallumon (7.2 táblázatban piros háttérrel jelölt mérések). Olyan eset is előfordulhat, amikor a predikció és a tényleges fogyasztás a megengedett intervallum kívül esik, azonban az egyik érték az intervallum egyik, a másik az intervallum másik végén esik kívül (7.2 táblázat első sora). Ilyenkor lényegében logikailag megfelelő javítás történik (azonban egy jobb predikció esetén hamis detekciót eredményezne a szabály illesztés).

7.0.1. Modellparaméter-optimalizáció

A megoldás finomra hangolása során két paramétert állítottam. Az egyik paraméter arra vonatkozik, hogy ha egy mérés túl közel esik egy fogyasztás intervallumhoz, akkor arra ne végezzünk szabályillesztést. Erre azért van szükség, mert ha predikció nagyon közel áll az intervallumhoz, akkor könnyen lehet, hogy egy rossz fogyasztási intervallumba lett sorolva, és a szabályillesztés hamis detekciót fog eredményezni.

A másik paraméter a távolságfüggvény által megadott maximális értékre vonatkozik. A távolságfüggvény azt határozza meg, hogy egy szabály milyen közel esik kvalitatív értelemben az adott megfigyeléshez. Amennyiben minden input paraméter a szabállyal azonos kategória besorolással rendelkezik, akkor a távolság (d) értéke nulla. Abban az esetben, ha az épület egy kategóriában tér el egy kvalitatív jellemzőt vizsgálva, akkor a távolság egy (d=1) lesz, kettő eltérés esetén pedig kettő lesz atávolság, és így tovább.

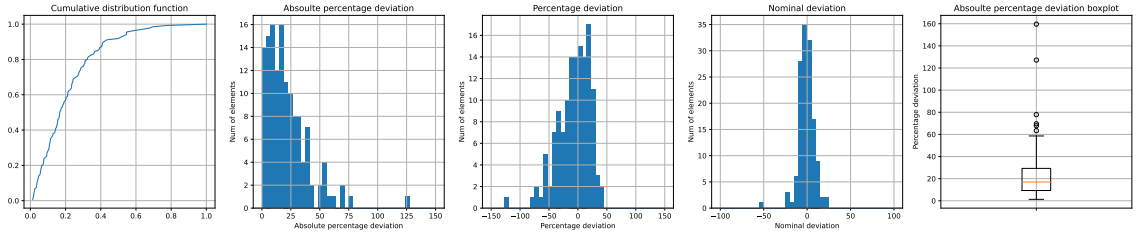
A távolságfüggvény növelésével nő a szabályrendszer lefedési képessége, azaz több mérésre tudunk szabályt illeszteni. Emellett a tesztadatok kiértékelése során bebizonyosodott, hogy minél nagyobb a megengedett maximális távolság a mérés és a szabály között, annál több hamis detekció keletkezik. A paraméter állítása egy kompromisszum alapú döntés, mennyi hamis detekciót engedünk meg a nagyobb lefedettség érdekében.

7.0.2. A megoldás kiértékelése a tesztadatokon

A tesztadatok kiértékelésekor a 5.2 alfejezetben bemutatott magyar modell becsléseit vettem referencia értéknek. A szabály illesztést kiértékeltem d $\leq$ 1, és d $\leq$ 2 távolságra (d $\leq$ 3 esetén már csak hamis detekciók keletkeznek). A vizsgálat során először mind a maximum, mind a minimum értékek illeszkedését ellenőriztem. Ezt követően csak a amximum túllépésre szűrtem rá. A tesztadatok vizsgálata során a kategória intervallum értékek és apredikció közötti százalékos eltérést 11%-ra állítottam (ezt az értéket használtam végül a validációs adatok kiértékelésre is).

7.0.2.1. Minimum és maximum értékek ellenőrzése

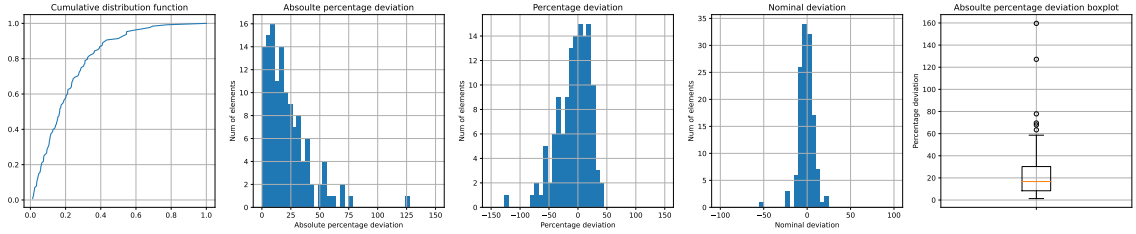
A kiugró értékek keresését elsőként d $\leq$ 1 távolságra végeztem el, az eredmény a 7.1 ábrán látható. A 5.1 ábrán bemutatott magyar modell kiértékeléséhez képest látható, hogy a MAPE 4%-al csökkent. A legfontosabb eltérés az, hogy a harmadik diagramokat összehasonlítva a negatív irányú kiugró eltérések nagy része le lett vágva. Az kiugró értékek levágása a boxplotokon is megfigyelhetőek. A negatív előjelű hibákat tekintjük minőségileg rosszabb hibáknak, ezért ez egy minőségileg kifejezetten jó kiugró érték detektálás.



{'cdf_score': 0.78907, 'MAPE': 22.67426, 'RMSE': 9.28691, 'MAE': 6.63058}

7.1. ábra. Minimum és maximum értékek ellenőrzése d $\leq$ 1 távolságra

A 7.2 ábrán látható a d $\leq$ 2 kiugró érték detekció eredménye. Megfigyelhető, hogy túlzott eltérés nincs a d $\leq$ 1 eredményéhez képest, a MAPE értékek is szinte megegyeznek.



{'cdf_score': 0.79011, 'MAPE': 22.65674, 'RMSE': 9.11761, 'MAE': 6.38225}

7.2. ábra. Minimum és maximum értékek ellenőrzése d $\leq$ 2 távolságra

A 7.4 táblázatban már jobban látszódik a két távolságérték által végzett detekció különbsége. Látható, hogy a d $\leq$ 1 egy, míg a d $\leq$ 2 kettőt darab jó mérést vág le. A d $\leq$ 2 ezenkívül levág még néhány nagyobb hibával rendelkező mérést is. Az eredményke valós kiértékeléséhez azonban elengedhetetlen, hogy megvizsgáljuk a hamis és jó detekciók arányát.

Hiba-intervallum	**0.0 - 0.15**	**0.15 - 0.2**	**0.2 - 0.266**	**0.266 - 0.3**	**0.3 - 0.5**	**0.5 - 0.7**	**0.7 - 0.9**	**0.9 - inf.**
Hun model	58	20	21	7	23	10	3	7
Outlier det. (d <=1)	57	19	19	7	21	10	1	2
Outlier det. (d <=2)	56	17	16	7	20	10	1	2

7.1. táblázat. Minimum és maximum érték ellenőrzés hiba intervallumai

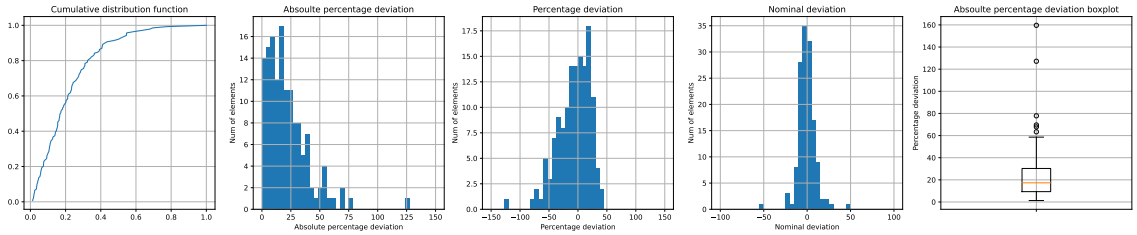
A konkrét szabály illesztéseket a 7.2 táblázat mutatja be. Piros háttérrel vannak jelölve azok a hamis detekciók, ahol logikailag rossz detektálás történt. Látható, hogy a két távolság megengedése két hamis és két jó predikciót eredményezett. A táblázat *TYPE* oszlopa tartalmazza, hogy az adott predikció a minimum vagy a maximum értékhatárt sértette-e meg. A táblázat legfontosabb észrevétele, hogy a minimum értékek detektálása csak két esetben hozott logikailag helyes vágást.

CONS MIN	**CONS MAX**	**CONS NUM**	**PRED NUM**	**SUM DIFF**	**ERROR**	**TYPE**
2	2	1	3	0	0.934	MAX
2	3	1	1	1	0.473	MIN
3	3	1	2	2	0.406	MIN
2	2	1	1	1	0.26	MIN
2	3	1	1	2	0.156	MIN
2	2	3	3	1	0.098	MAX
1	1	1	2	1	1.4	MAX
1	2	1	3	1	1.131	MAX
1	1	1	2	1	1.07	MAX
1	1	1	3	1	1.05	MAX
1	1	1	2	1	0.851	MAX
1	1	1	2	1	0.734	MAX
4	5	4	3	0	0.345	MIN
3	3	3	2	2	0.265	MIN
1	1	1	2	2	0.240	MAX

7.2. táblázat. Minimum és maximum érték ellenőrzés szabály illesztései

7.0.2.2. Maximum értékek ellenőrzése

Az előző alfejezet eredményeiből kifolyólag a kiugró érték ellenőrzést elvégeztem csak a maximum értékek vizsgálatára is ($d \leq 2$ távolságra). A 7.3 ábrán látható, hogy a MAPE érték lényegileg nem változott, és a diagramok is hasonló képet mutatnak (ha nagyon megnézzük felfedezhető, hogy – az előzőekben rossz logikai úton – levágott 40% körüli hibával rendelkező predikciók megjelentek a hiba eloszlásban).



7.3. ábra. Maximum értékek ellenőrzése d $\leq$ 2 távolságra

A 7.4 táblázatban látható, hogy a maximum érkek ellenőrzése utáni kiugró érték detekció hiba-intervallum táblázata. A minimum érték detekciót elhagyva egy logailag helytálló kiugró érték detekciót kapunk, ahol lényegében egy hamis detekció fordul elő, miközben a 70% feletti hibával rendelkező predikciók csupán 30%-a marad meg. Amennyiben d $\leq$ 2 távolságot választjuk, eggyel több kiugró értéket szűrünk ki (0.266 - 0.3 hiba intervallumban).

Hiba-intervallum	**0.0 - 0.15**	**0.15 - 0.2**	**0.2 - 0.266**	**0.266 - 0.3**	**0.3 - 0.5**	**0.5 - 0.7**	**0.7 - 0.9**	**0.9 - inf.**
Hun model	58	20	21	7	23	10	3	7
Outlier det. (d <=1)	57	20	21	7	23	10	1	2
Outlier det. (d <=2)	57	20	20	7	23	10	1	2

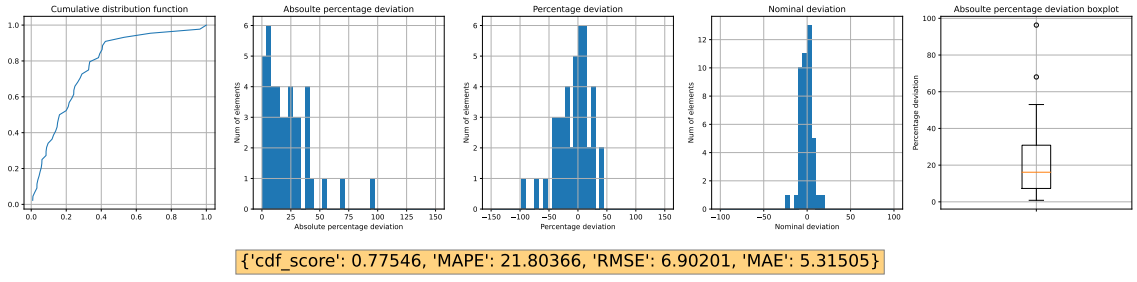
7.3. táblázat. Maximum értékek ellenőrzés hibaintervallumai

7.0.3. A megoldás kiértékelése a validációs adatokon

A validációs adatok kiértékelésekor a 5.4.3 alfejezetben bemutatott magyar modell predikcióit szolgálnak referencia értékként. A tesztaadtokon elvégzett kiugró érték ellenőrzés alapján a validációs adatokon csak maximum ellenőrzést végeztem d $\leq$ 2 távolságra.

A 7.4 ábrán látható a validációs adatokon végzett kiugró érték szűrés eredménye. A 5.5 eredménnyel összehasonlítva látható, hogy a szűrés a vártnál még jobban teljesített (a véletlen mintavételezés során arányaiban több nagyobb hibával rendelkező mérés került a validáció adatszetbe, mint a tesztkészletbe). A MAPE értéken 6%-ot sikerült csökkenteni. Látható, hogy a pozitív hibával rendelkező mérések nagy része eltűnt.

A 7.4 táblázatot megnézve látható, hogy a 90% feletti hibák háromnegyed, az 50% feletti hibák több. Mint 60%-a kiszűrésre került, miközben csupán egy alacsony hibával rendelkező megfigyelés lett eltávolítva.



7.4. ábra. Maximum értékek ellenőrzése d $\leq$ 2 távolságra

A 7.5 táblázat mutatja be a konkrét szabály illesztéseket a validációs adathalmazon. Látható, hogy egyetlen esetben történt logikailag nem megfelelő detekció. A másik hat esetben pedig sikerült kiugróan nagy hibákat is kiszűrni.

Hiba-intervallum	**0.0 - 0.15**	**0.15 - 0.2**	**0.2 - 0.266**	**0.266 - 0.3**	**0.3 - 0.5**	**0.5 - 0.7**	**0.7 - 0.9**	**0.9 - inf.**
Hun model	20	4	7	3	8	4	0	4
Outlier det. (d <=2)	19	4	7	2	8	2	0	1

7.4. táblázat. Maximum értékek ellenőrzés hibaintervallumai

A 7.5 táblázatot megnézve látható, hogy a 90% feletti hibák háromnegyed, az 50% feletti hibák több, mint 60%-a kiszűrésre került, miközben csupán egy alacsony hibával rendelkező megfigyelés lett eltávolítva.

CONS MIN	**CONS MAX**	**CONS NUM**	**PRED NUM**	**SUM DIFF**	**ERROR**
3	3	4	4	2	0.014
1	3	3	4	2	1.21
1	1	1	2	2	0.929
1	1	1	2	2	0.904
1	1	1	2	1	0.684
1	1	1	2	1	0.534
3	3	3	4	1	0.289

7.5. táblázat. Maximum érték ellenőrzés szabály illesztései

8. fejezet

Összefoglalás

Dolgozatomban ismertettem egy új megközelítést kvalitatív transzfer-tanítás alkalmazására. A javasolt megközelítés mind épület-benchamarking, mind transzfer-tanítás szempontjából újszerűnek tekinthető. A dolgozatban bemutatott elméletet egy gyakorlati alkalmazásba is átültettem. A munkám során a következő főbb eredményeket értem el:

- **Feltáró és összehasonlító adatelemzés a UK és magyar adatbázison:** Munkámban megvizsgáltam és összehasonlítottam a UK és magyar adatkészlet eloszlásait, korrelációit. Főkomponens analízist végeztem mindkét adathalmazon, és összevetettem az eredményeket. A feltáró adatelemzés megmutatta, hogy egy transzfer-tanítási logikailag megalapozott.

- **Épületek kvalitatív modellezése energiahatékonyság szempontjából:** A kutatásom során létrehoztam egy kvalitatív modellezési paradigmát valós épületek energiahatékonyságának jellemzésére. A közös kvalitatív modell segítségével a különböző országok adatbázisai összhasonlíthatóvá és leképezhetővé válnak.

- **Kvalitatív-kvantitatív transzfer-tanítás alkalmazása relációs tudásátvitelre:** Dolgozatomban egy új megközelítést mutattam be a reláció alapú transzfer-tanításra. A megközelítés lényege, hogy azonos logikai struktúrával rendelkező adatkészletek kvalitatív modellezésével átvihetőek az alapvető szabályszerűségek az egyes domének között.

- **Kiugró értékek keresése transzfer-tanítással létrehozott szabályrendszerrel:** Javasoltam egy megoldást kiugró értékek keresése a dolgozatban ismertetett transzfer-tanítási megközelítés segítségével. A megközelítés alapgondolata, hogy a kategorikus értelembe vett terjedelmesebb adatbázisból átvett szabályok alkalmasak meglévő modellek felülvezérlésére.

- **Ellenőrző-szabályrendszer kiterjesztése kényszerprogramozás segítségével:** A dolgozatban ismertettem egy megoldást az állapottert csupán részlegesen lefedő szabályrendszerek kiterjesztésére. A javasolt megoldás lényege hogy kényszerprogramozás segítségével megkeressük az adott megfigyeléshez legközelebb eső szabályt. A kényszerprogramozásank hála könnyedén alkalmazhatunk további kvalitatív következtetési módszereket.

- **A megvalósított megközelítés gyakorlati alkalmazása valós hazai adatokon:** Az javasolt megoldást egy valós problémán szemléltettem és validáltam. A tesztadatok segítségével a kiugró érték ellenőrző alkalmazást finomra hangoltam. Ezt követően egy ötven megfigyelésből álló validáló készleten alkalmaztam a megoldásomat. A kiugró érték detektálással a 90% feletti hibák háromnegyed, az 50%

feletti hibák több, mint 60%-a kiszűrésre került, miközben csupán egy hamis detekció keletkezett.

A munkámat az új elméleti megközelítéseken kívül valós gyakorlati eredményekkel zártam. A megközelítés azonban számos helyen tovább kutatható és fejleszthető. A főbb továbbfejlesztési irányok a következők:

- **Transzfer leképezés további optimalizálása:** A jelenlegi output transzfer-leképezés még nem minden esetben hibátlan. Bizonyos kategória-kombinációkra nézve eltérő HUN-UK transzer output értékeket kapunk. Ebből kifolyólag a transzfer leképezés további optimalizálása egy továbbfejlesztési aspektusa a munkámnak.
- **Minimum érték detekció felülvizsgálata:** A jelenlegi kiugró érték kereső megoldás a minimum értékek kiszűrésére nem teljesít jól, így a munkámban csak a maximum értékeket vizsgáltam. A minimum érték detekció hibáját a későbbi munkámban felül kívánom vizsgálni.
- **K-legközelebbi szabálykeresés további optimalizációja:** Továbbfejlesztési opcióként tovább szeretném optimalizálni a k-legközelebbi szabálykeresést például súlyozás alkalmazásával.
- **További kvalitatív következtetési módszerek alkalmazása:** Az eredményekből kiderült, hogy túl nagy távolságokra nem működik a szabálykeresés. A szabályrendszert ezért tovább kívánom bővíteni, hogy a teljes állapotteret le tudjam fedni szabályokkal. Ahol nincs elegendő mérés, ott valamilyen kvalitatív következtetés útján kívánok szabályokat alkotni. Továbbá a manuális – szakértői – szabály inputálást, illetve a *latice* alapú megközelítést is integrálni kívánom a megoldásban.
- **Esetleges folytonos transzfer vizsgálata a kvalitatív modellek segítségével:** Továbbfejlesztési opcióként tartom számon, hogy a jelenlegi kvalitatív kategóriale-képezést egy folytonos leképezésre cseréljem le, a kvalitatív modellek megtartása mellett. Az edigi folytonos transzfer megközelítésem nem kvalitatív modellezésen alpszik, ezért ez egy új megközelítés lenne.

Irodalomjegyzék

[1] Aleksandra Arcipowska – Filippos Anagnostopoulos – Francesco Mariottini – Sara Kunkel: *Energy performance certificates across the EU*. 2014. október, Buildings Performance Institute Europe (BPIE). ISBN 9789491143106.

[2] Antonio Attanasio – Marco Piscitelli – Silvia Chiusano – Alfonso Capozzoli – Tania Cerquitelli: Towards an Automated, Fast and Interpretable Estimation Model of Heating Energy Demand: A Data-Driven Approach Exploiting Building Energy Certificates. *Energies*, 12. évf. (2019), 1273. p.

[3] BRE: England and wales historic fuel prices (2022.05.26.). `https://files.bregroup.com/SAP/PCDB%20fuel%20prices%20January%202022.xlsx`.

[4] Bert Bredeweg – Floris Linnebank – Anders Bouwer – Jochem Liem: Garp3 — workbench for qualitative modelling and simulation. *Ecological Informatics*, 4. évf. (2009) 5. sz., 263–281. p. ISSN 1574-9541. URL `https://www.sciencedirect.com/science/article/pii/S1574954109000818`. Special Issue: Qualitative models of ecological systems.

[5] FRANCESCO CALIMERI – GIOVAMBATTISTA IANNI – FRANCESCO RICCA: The third open answer set programming competition. *Theory and Practice of Logic Programming*, 14. évf. (2012. sep) 1. sz., 117–135. p.
URL `https://doi.org/10.1017%2Fs1471068412000105`.

[6] Sonia Williams (Data Science Campus): Using machine learning to predict energy efficiency (2022.10.13.). `https://datasciencecampus.ons.gov.uk/projects/using-machine-learning-to-predict-energy-efficiency/`.

[7] Tianqi Chen – Carlos Guestrin: XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16 konferenciasorozat. New York, NY, USA, 2016, ACM, 785–794. p. ISBN 978-1-4503-4232-2.
URL `http://doi.acm.org/10.1145/2939672.2939785`. 10 p.

[8] Francois Chollet és mások: Keras, 2015.
URL `https://github.com/fchollet/keras`.

[9] Xuefeng Gao – Ali Malkawi: A new methodology for building energy performance benchmarking: An approach based on intelligent clustering algorithm. *Energy and Buildings*, 84. évf. (2014. december), 607–616. p. ISSN 0378-7788. URL `https://www.sciencedirect.com/science/article/pii/S0378778814006720`.

[10] Ian T. Jolliffe – Jorge Cadima: Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374. évf. (2016) 2065. sz., 20150202. p. URL `https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2015.0202`.

[11] Guolin Ke – Qi Meng – Thomas Finley – Taifeng Wang – Wei Chen – Weidong Ma – Qiwei Ye – Tie-Yan Liu: Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30. évf. (2017), 3146–3154. p.

[12] Imre Kocsis: Qualitative models in resilience assurance. 2019.
URL https://repozitorium.omikk.bme.hu/handle/10890/13122. PhD dissertation.

[13] Benjamin Kuipers: Qualitative reasoning: Modeling and simulation with incomplete knowledge. *Automatica*, 25. évf. (1989) 4. sz., 571–585. p. ISSN 0005-1098. URL https://www.sciencedirect.com/science/article/pii/000510988990099X.

[14] Lilyana Mihalkova – Tuyen Huynh – Raymond J. Mooney: Mapping and revising markov logic networks for transfer learning. In *Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 1*, AAAI'07 konferenciasorozat. 2007, AAAI Press, 608–614. p. ISBN 9781577353232. 7 p.

[15] Baumann Mihály – Dr. Csoknyai Tamás – Dr. Kalmár Ferenc – Dr. Magyar Zoltán – Dr. Majoros András – Dr. Osztroluczky Miklós – Szalay Zsuzsa – Prof. Zöld András: *Épületenergetika (Segédlet)*. 2009. április, PTE Pollack Mihály Műszaki Kar. ISBN 978-963-7298-31-8.

[16] Netjogtár: 7/2006. (v. 24.) tnm rendelet az épületek energetikai jellemzőinek meghatározásáról. https://net.jogtar.hu/getpdf?docid=a0600007.tnm&targetdate=20210101&printTitle=7/2006.+%28V.+24.%29+TNM+rendelet.

[17] SAJN Nikolina: Energy efficiency of buildings: A nearly zero-energy future? https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2016)582022.

[18] OpenDataCommunities: Energy performance of buildings data england and wales (guidance). https://epc.opendatacommunities.org/docs/guidance.

[19] Google OR-Tools: Constraint optimization (2022.10.28.). https://developers.google.com/optimization/cp.

[20] Oleksii Pasichnyi – Jörgen Wallin – Fabian Levihn – Hossein Shahrokni – Olga Kordas: Energy performance certificates — new opportunities for data-enabled urban energy policy instruments? *Energy Policy*, 127. évf. (2019. 01), 486–499. p.

[21] Liudmila Prokhorenkova – Gleb Gusev – Aleksandr Vorobev – Anna Veronika Dorogush – Andrey Gulin: Catboost: unbiased boosting with categorical features, 2017.
URL https://arxiv.org/abs/1706.09516.

[22] Matthew Richardson – Pedro Domingos: Markov logic networks. *Machine Learning*, 62. évf. (2006. február) 1. sz., 107–136. p. ISSN 1573-0565.
URL https://doi.org/10.1007/s10994-006-5833-1.

[23] Francesca Rossi – Peter van Beek – Toby Walsh: *Handbook of Constraint Programming*. USA, 2006, Elsevier Science Inc. ISBN 9780080463803.

[24] Stuart Russell – Peter Norvig: *Artificial Intelligence: A Modern Approach*. 3. kiad. 2010, Prentice Hall.

[25] Lisa A. Torrey – J. Shavlik: Chapter 11 Transfer Learning. *undefined*, 2009. URL https://www.semanticscholar.org/paper/Chapter-11-Transfer-Learning-Torrey-Shavlik/1890c124749d00cce965e0b9495eafe127e16a26.

[26] Qiang Yang – Yu Zhang – Wenyuan Dai – Sinno Pan: *Transfer Learning*. 2020. január, Cambridge University Press. ISBN 978-1-107-01690-3.

[27] Fuzhen Zhuang – Zhiyuan Qi – Keyu Duan – Dongbo Xi – Yongchun Zhu – Hengshu Zhu – Hui Xiong – Qing He: A comprehensive survey on transfer learning, 2019. URL https://arxiv.org/abs/1911.02685.

[28] Pannon Építőműhely Kft.: U érték definíció (2022.10.13.). https://www.pannonmuhely.hu/energetika/hoszigeteles-szotar.php#U-ertek.

Függelék

F.1. Optimális szabálykeresés kényszerprogramozással

```
def find_optimal_building(hun_walls, hun_roof, hun_mainheat, hun_windows, hun_mainheat_c,
                                                hun_hotwater, hun_age, hun_area, hun_pred):

    model = cp_model.CpModel()

    # Szabaly indexet tarolo valtozo
    index = model.NewIntVar(0, num_of_uk_houses -1, 'index')

    # Fal jellemzo (szabaly - megfigyeles) ekvivalencia kikenyszeritese
    walls_val = model.NewIntVar(1, 5, 'walls_val')
    model.AddAllowedAssignments([index,walls_val], walls_tuple)
    model.AddAbsEquality(hun_walls, walls_val)

    # Teto jellemzo (szabaly - megfigyeles) kulonbseg valtozo letrehozasa
    roof_val = model.NewIntVar(1, 5, 'roof_val')
    model.AddAllowedAssignments([index,roof_val], roof_tuple)
    roof_diff_abs = model.NewIntVar(0, 4, 'roof_diff_abs')
    model.AddAbsEquality(roof_diff_abs, hun_roof- roof_val)

    # Futes jellemzo (szabaly - megfigyeles) kulonbseg valtozo letrehozasa
    mainheat_val = model.NewIntVar(1, 5, 'mainheat_val')
    model.AddAllowedAssignments([index,mainheat_val], mainheat_tuple)
    mainheat_diff_abs = model.NewIntVar(0, 4, 'mainheat_diff_abs')
    model.AddAbsEquality(mainheat_diff_abs, hun_mainheat- mainheat_val)

    # Ablakok jellemzo (szabaly - megfigyeles) kulonbseg valtozo letrehozasa
    windows_val = model.NewIntVar(1, 5, 'windows_val')
    model.AddAllowedAssignments([index,windows_val], windows_tuple)
    windows_diff_abs = model.NewIntVar(0, 4, 'windows_diff_abs')
    model.AddAbsEquality(windows_diff_abs, hun_windows-windows_val)

    # Futes vezerles jellemzo (szabaly - megfigyeles) kulonbseg valtozo letrehozasa
    mainheat_c_val = model.NewIntVar(1, 5, 'mainheat_c_val')
    model.AddAllowedAssignments([index,mainheat_c_val], mainheat_c_tuple)
    mainheat_c_diff_abs = model.NewIntVar(0, 4, 'mainheat_c_diff_abs')
    model.AddAbsEquality(mainheat_c_diff_abs, hun_mainheat_c-mainheat_c_val)

    # Melegviz jellemzo (szabaly - megfigyeles) kulonbseg valtozo letrehozasa
    hotwater_val = model.NewIntVar(1, 5, 'hotwater_val')
    model.AddAllowedAssignments([index,hotwater_val], hotwater_tuple)
    hotwater_diff_abs = model.NewIntVar(0, 4, 'hotwater_diff_abs')
    model.AddAbsEquality(hotwater_diff_abs, hun_hotwater-hotwater_val)
```

F.1. ábra. Optimális szabálykeresés kényszerprogramozással (1)

```
    # Epites eve jellemzo (szabaly - megfigyeles) kulonbseg valtozo letrehozasa
    age_val = model.NewIntVar(1, 5, 'age_val')
    model.AddAllowedAssignments([index,age_val], age_tuple)
    age_diff_abs = model.NewIntVar(0, 4, 'age_diff_abs')
    model.AddAbsEquality(age_diff_abs, hun_age-age_val)

    # Alapterulet jellemzo (szabaly - megfigyeles) kulonbseg valtozo letrehozasa
    area_val = model.NewIntVar(1, 5, 'area_val')
    model.AddAllowedAssignments([index,area_val], area_tuple)
    area_diff_abs = model.NewIntVar(0, 4, 'area_diff_abs')
    model.AddAbsEquality(area_diff_abs, hun_area-area_val)

    # Szabaly serules kikenyszeritese
    cons_max_val = model.NewIntVar(1, 5, 'cons_max_val')
    model.AddAllowedAssignments([index,cons_max_val], cons_max_tuple)

    cons_min_val = model.NewIntVar(1, 5, 'cons_min_val')
    model.AddAllowedAssignments([index,cons_min_val], cons_min_tuple)

    greater_than_max = model.NewBoolVar("greater_than_max")
    lower_than_min = model.NewBoolVar("lower_than_min")

    hun_pred_val = model.NewIntVar(hun_pred, hun_pred, 'hun_pred_val')

    model.Add(hun_pred_val > cons_max_val).OnlyEnforceIf(greater_than_max)
    model.Add(hun_pred_val < cons_min_val).OnlyEnforceIf(lower_than_min)

    model.AddBoolOr([greater_than_max, lower_than_min])

    # Optimalizacios (minimalizalasi) feltetel
    model.Minimize(roof_diff_abs +
                   mainheat_diff_abs +
                   windows_diff_abs +
                   mainheat_c_diff_abs +
                   hotwater_diff_abs +
                   age_diff_abs+
                   area_diff_abs)

    # A problema megoldasa
    solver = cp_model.CpSolver()
    status = solver.Solve(model)

    # Amennyiben a problema megoldhato
    if status == cp_model.OPTIMAL or status == cp_model.FEASIBLE:

        sum_diff = solver.ObjectiveValue()
        index = solver.Value(index)

        # Visszateres a szabaly indexevel es az ossztavolsaggal
        return [index, sum_diff]

    else:
        print('No solution found.')
        return np.NaN
```

F.2. ábra. Optimális szabálykeresés kényszerprogramozással (2)