MÁDER MIKLÓS PÉTER[1]

ADATHIÁNY KEZELÉSI ELJÁRÁSOK ÉS IMPUTÁLÁSOK ÖSSZEHASONLÍTÁSA[2]

A KUTATÁS FÓKUSZA

Az értekezés témája az adathiányok jellegzetességeinek bemutatása és a különböző adathiány kezelő eljárások összehasonlítása, értékelése. A téma választását az indokolja, hogy az adathiány szinte minden adatgyűjtés során megjelenik. Az adathiányok három típus szerint rendezhetők, melyek közül kettő szisztematikusan torzítja a megfigyelhető eloszlást. Ezért a jelen értekezés a szakirodalomnak az adathiány problematikájára adott válaszait ezen három adathiány típus mentén vizsgálja, oly módon, hogy ellenőrizhető körülmények között modellezi az adathiány-kezelő eljárásokat, majd értékeli az eljárások eredményeit.

A kutatók az általános gyakorlat szerint, ha a válaszolókra vonatkozó megfigyelések hiányosak, lehetőleg kihagyják az adathiányos eseteket a statisztikai elemzésekből. Az adathiányos esetek figyelmen kívül hagyása helyett az adathiányok okainak és mintázatainak feltérképezése után az adathiányok pótlása, az úgynevezett imputálás válhat az adathiány kezelés főáramává. Az imputálás eszközét az elemzés során akkor használják, ha az adott hiányzó értéket becslésen alapuló értékkel helyettesítik (Rubin 1987). Az imputálás az a folyamat, amely az adatbázisban lévő adathiányokat egy becslési algoritmus segítségével feltölti. Az értekezésben bemutatott szakirodalom szerint számos alkalommal a megfigyelt esetek információtartalma hordoz annyi tudást önmagában, hogy az adathiányok feltöltésére nézve egy matematikailag releváns becslést lehessen adni.

Számos, a komplexitásában jelentős mértékben különböző megoldás létezik az adathiányok becslésére nézve. Az értekezés a különböző adathiány-kezelő eljárások bemutatása és modellezése során a *Handbook of Statistical Modeling for the Social and Behavioral Sciences* J.A.R. Little és N, Schenker által írt fejezetekre (Little and Schenker 1995), valamint D.B. Rubin *Multiple Imputation for Nonresponse in Surveys* című szakkönyvére támaszkodik (Rubin, 1987). A szakirodalom számos eljárást említ, a feltételhez nem kötött átlag behelyettesítéstől a többszörös imputálásig, ám ezek közül az eljárások közül nem mindegyik vezet megfelelő eredményre.

A KUTATÁS CÉLJA

Az értekezés során alkalmazott munkahipotézis alapja az a feltételezés, hogy az imputálási eljárások valóban csökkentik a hiányos adatbázisok által adott becslések torzítottságát. Ezen hipotézis ellenőrzésére az értekezés a három különböző adathiány-mechanizmus mentén modellezi az adathiány-kezelő eljárások (a különböző missingelési típusok) és a különböző imputálási eljárások hatékonyságát.

A DISSZERTÁCIÓ ÁTTEKINTÉSE

Az értekezés először bemutatja az adathiányok esetlegesen megfigyelhető torzító hatását (1). Majd számba veszi az adathiányok jellegzetességeit, mechanizmusait. Áttekintést nyújt az imputálási módszerekről, az úgynevezett naivaktól kezdve az összetettebb módszereken át, a többszörös imputálási eljárásokig (2). Ezt követően – a fent már említett – három különböző adathiány-mechanizmus mentén modellezi az imputálási eljárásokat (3). Végül összehasonlítja a különböző eljárások hatékonyságát (4).

Az adathiányok torzító hatása

Az adathiányok esetében fontos szempont, hogy – a számos lehetséges közül – milyen okra vezethetők vissza, illetve hogy mekkora az adathiány mértéke. Az 1% vagy ez alatti adathiány ráta McDermit szerint triviális, az 1-5% közötti kezelhető. Az 5-15% közötti adathiány kezelése már szofisztikált módszerek használatát igényli, 15% feletti adathiány pedig már súlyos interpretálási problémákat vet fel (McDermit 1999).

Az adathiány torzító hatásának matematikai hátterét vázolja fel Rudas Tamás a *Mixture Models of Missing Data* című cikkében (Rudas 2005). Elméleti megközelítésében a vizsgált populáció két részre bontható: az egyikbe tartoznak azok, akik egy

[1] Eötvös Lóránd Tudományegyetem Társadalomtudományi Kar, Szociológia Doktori Iskola
[2] Doktori értekezés tézisei, ELTE TáTK Szociológia Doktori Iskola, Budapest 2012, témavezető: Dr. Székelyi Mária DSc.

kérdőíves kutatás számára elérhetőek, és emiatt vizsgálhatók, míg a másik csoportban olyanok vannak, akik nem elérhetőek, vagyis nem is vizsgálhatók.

A teljes populációra becsült eloszlás a két csoporthoz tartozó eloszlások együttese, vagyis a megfigyelhető (*M*) és a nem-megfigyelhetőé (*N*):

$$(1-p)M+pN,$$

ahol az 1-p paraméter jelöli a megfigyelhetőség szempontjából elérhető esetek arányát a populációban, p pedig a megfigyelhetőség szempontjából nem-elérhető esetek arányát. Ha a p értéke nulla, akkor sikerült a teljes populációból torzítatlanul mintát venni, és azt lekérdezni.

A survey kutatásra vonatkoztatott elérhetőség és nem-elérhetőség dichotómiája helyett helyesebb azt a megközelítést alkalmazni, mely szerint a populáció minden egyes egyedére kalkulálható egy survey kutatásban való részvételt becslő pontos valószínűség. Azokra, akikre ez a survey kutatásban való részvételre vonatkozó valószínűségi becslés nulla, azok a nem-megfigyelhetők csoportjába kerülnek. A nem-megfigyelhetők csoportjába tartozók esetében a teljes megfigyelési adatdokumentáció hiányzik, ez az úgynevezett *Unit nonresponse*.

Az adathiányok jellegzetességei

- Az MCAR adathiány-mechanizmus lényege, hogy az adathiány nem függ ellenőrző változóktól. Az adathiány feltehetően egy előre tervezett módszertani szempont miatt keletkezett.
- Az MAR adathiány-mechanizmus esetén egy változóban megfigyelhető adathiány más változók függvénye, azokkal kapcsolatban van, azaz azokkal statisztikai összefüggés mutatható ki.

A NOTMAR adathiány-mechanizmus esetében egy adott változón belüli adathiány önmagának, a változónak a függvénye. Ebben az esetben, az adatmátrixban lévő változók összefüggése, illetve függetlensége az adathiányos változóval már nem releváns (Rubin 1976).

Az imputálási eljárások modellezése

Az imputálási módszereket Laaksonen négy fő kategóriába osztotta (Laaksonen 1999), amelyek közül az első valójában nem is a szó szoros értemében vett imputálási eljárás, de mégis egyfajta adathiány kezelési módszer.

1. A CC és AC eljárások, ahol az adathiány értékeket nem imputálják. Az elemzés során az adathiányos eseteket általában kihagyják az elemzésből. Ha csak a teljes, komplett eseteket elemzik, az a *Complete-Case analysis* (CC), ha pedig csak az adott összefüggés, asszociáció vizsgálatára vonatkozóan hagyják figyelem kívül az adathiányos eseteket akkor az *Available Cases analysis* (AC) módszert alkalmazzák.
2. Deduktív vagy logikai imputálás, ahol egy ismert, jogosan létező adathiány miatt logikailag imputálhatók az adathiányos esetek. Ilyenre példa a kérdés válaszopcióihoz rendelt ugrás a kérdőív kérdéssorrendjében.
3. Az imputált adathiányok értékei egy modell eredményeként születnek, emiatt lehetséges, hogy a megfigyelt esetek között nincs az újonnan imputált értékhez hasonló, annak megfeleltethető érték. Ezt Laaksonen model-donor imputálásnak (*model-donor imputation*) nevezte.
4. Az imputálás alapjául már megfigyelt esetek értékei szolgálnak. Ez a valódi donor imputálás (*real-donor imputation*). A harmadik csoport szerinti imputálás is valós értelmezési tartományból generál imputált értékeket, de az az érték nem a már megfigyelt esetek közül kerül kiválasztásra, nem tényleges megfigyelésen alapul, mint az a negyedik módszer esetében.

Az imputálási eljárások által kapott eredményeket először négy, logikailag különálló csoportban ábrázoljuk:

1. Az első csoportba került eljárások a legegyszerűbb megközelítések közé tartoznak, ezek a CC, AC, és a súlyozás. Ezek az esetek még nem imputálások, de mégis a hiányzó adatok ismeretében végzett eljárásmódok (Laaksonen 2000). A csoportban használt rövidítések:
   - *Complete Cases*, vagyis a *listwise deletion*: CC
   - *Available Cases*, vagyis a *pairwise deletion*: AC
   - Súlyozás: W

2. Már az imputálási eljárások közé tartozik az átlag, medián, és a módusz behelyettesítésének technikája; ez a második csoport. A csoportban használt rövidítések:
   - Átlag imputálás: MEAN
   - Módusz imputálás: MODUS
   - Medián imputálás: MEDIAN
3. A harmadik csoportba tartoznak azok az imputálási módszerek, amelyek az adathiányok helyére egy explicit modell alapján becsült értéket illesztenek. Ilyenek a regressziós imputálás, a reziduálisokkal bővített regressziós imputálás, a reziduálisokkal bővített regressziós imputáláson alapuló többszörös imputlálás (*multiple imputation*), és az elvárás maximalizáló (*expectation-maximization*) E.M. eljárás. A csoportban használt rövidítések:
   - Regressziós imputálás: REG
   - Reziduálisokkal bővített regressziós imputálás: REG+REZ
   - Reziduálisokkal bővített regressziós imputáláson alapuló többszörös imputálás: MI. A REG+REZ tudja biztosítani azt a következetes, mégis random algoritmust, melyet egységesen tudunk alkalmazni az MI esetében. Itt a REG imputálás azért nem lenne hatékony, mert minden imputálási eljárásában ugyanaz lenne az adott becslés értéke.
   - EM algoritmus: EM. Ennek alapja a REG imputálások iterálsa. elynek során az iterálási folyamat sikerét egy következetesen nem randomizált eljárással biztosítják.

   A negyedik csoportba tartozik a valós, ténylegesen megfigyelt donorok adta imputálási módszer. Ez implicit modell. Ebben a csoportban használt rövidítés:
   - *Real donor hot deck* imputálás: HOT DECK

A különböző eljárások hatékonyságának összehasonlítása

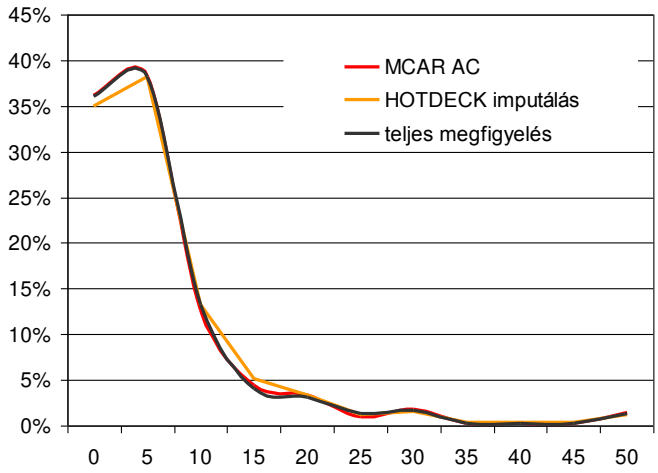
A modellezett imputálási eljárásokat 4 dimenzió mentén hasonlítjuk össze.

1. Először bemutatjuk az imputált változók viszonyát a magyarázó változókkal,
2. majd bemutatjuk a létrejött új változók és az eredeti változó eloszlásbeli különbségeit.
3. Ezt követően csak az imputált értékekre fordítjuk a figyelmet, megvizsgáljuk a kitörölt és a modellezés során a helyére imputált érték viszonyát,
4. végül egy *scoring* eljárással megpróbáljuk a számos különböző értékelő szempontot összefoglalni, és megnevezni a leghatékonyabbnak bizonyult imputálási eljárást.

Összegezve az MCAR adathiány–mechanizmus esetében a szofisztikált imputálási eljárásokat, megállapíthatjuk, hogy nem minden eljárás vezet torzítatlan becslésekhez. A törölt esetek átlagát, illetve heterogenitását nem minden eljárás imputálta helyesen. MCAR adathiány-mechanizmus esetében a HOT DECK eljárás bizonyult a leghatékonyabb imputálásnak, ez az eljárás a törölt esetek eloszlásának mind az átlagát, mind a szórását jól becsülte.

Végezetül egy összefoglaló ábrán igyekszünk illusztrálni, hogy az MCAR adathiány-mechanizmus modellezése során milyen volt az eredeti eloszlás, mi volt az adathiányok jellegzetessége, és milyen eredményre jutottunk az imputálási eljárások egyik jól teljesítő algoritmusával.

Az MCAR adathiány-mechanizmus és az imputálások hatása az adatbázisra.
Az imputált és a megfigyelt értékek eloszlása az adatbázisban

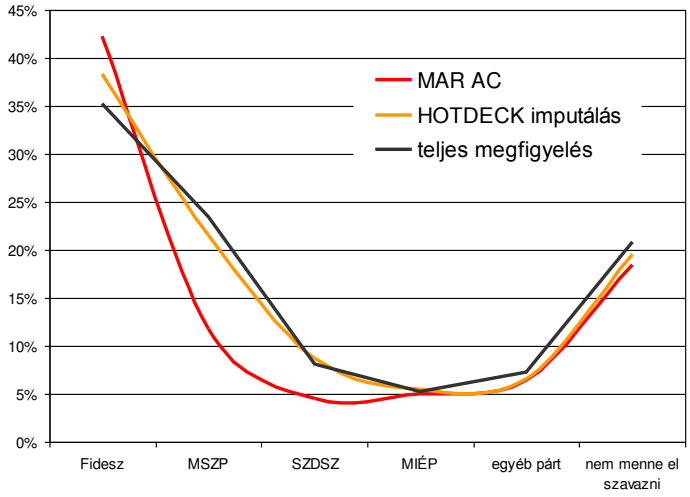


Jól látható, hogy az MCAR adathiány-mechanizmus nem jelentett igazán torzító tényezőt az eredetileg ismert változóra nézve, a HOT DECK imputálási eljárás pedig nagyon jól közelítette az eredeti eloszlást.

Összegezve az MAR adathiány–mechanizmus esetében a szofisztikált imputálási eljárásokat megállapíthatjuk, hogy nem minden eljárás vezet torzítatlan becslésekhez. A törölt értékek heterogenitását nem minden eljárás imputálta helyesen.

Végezetül egy összefoglaló ábrán igyekszünk illusztrálni, hogy az MAR adathiány-mechanizmus modellezés során milyen volt az eredeti eloszlás, mi volt az adathiányok jellegzetessége, és milyen eredményre jutottunk az imputálási eljárások egyik jól teljesítő algoritmusával.

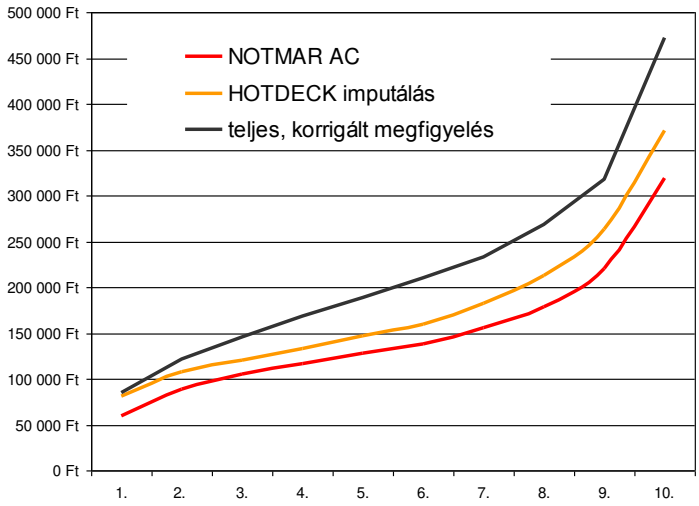
Az MAR adathiány-mechanizmus és az imputálások hatása az adatbázisra
Az imputált és a megfigyelt eloszlások a pártpreferencia változóban



Jól látható, hogy az MAR adathiány-mechanizmus esetében az AC határozottan torz eloszlást mutat az eredetileg ismert változóhoz viszonyítva, a HOT DECK imputálási eljárás viszont ezt a torzítást nagyon jól korrigálta és az AC eloszlást az eredeti eloszlás irányába módosította. Ez természetesen annak köszönhető, hogy az MAR adathiány-mechanizmus esetében az adathiány az adatbázisban megfigyelhető változók függvénye, emiatt az adatbázisban lévő információtartalom segítségével az MAR adathiány-mechanizmusú adathiány sikeresen imputálható.

A NOTMAR mechanizmusú adathiány esetében is egy összefoglaló ábrán igyekszünk illusztrálni, hogy ezen adathiány-mechanizmus modellezése során milyen volt az eredeti eloszlás, mi volt az adathiányok jellegzetessége, és milyen eredményre jutottunk az imputálási eljárások egyik jól teljesítő algoritmusával.

A NOTMAR adathiány-mechanizmus és az imputálások hatása az adatbázisra
A megfigyelt, a korrigált és az imputált jövedelmi átlagok decilisenként



A NOTMAR adathiány-mechanizmus esetében a megfigyelt esetek alkotta eloszlás nagymértékben alulbecsli a makro-statisztikák alapján feltételezhető eloszlást.

A korrigálás és a HOT DECK imputálás együttese sem volt képes olyannyira módosítani az adatfelvétel eloszlását, hogy az jól becsülje a makro-statisztikák által valószínűsített eloszlást.

Az imputálásokra nézve elmondható: annak ellenére, hogy az adatbázis alkotta megfigyelhető esetek *(M)* adta becslés közelített a teljes populáció paramétereihez, nem volt képes olyan magas értékek imputálására, amelyek jelentős mértékben korrigálni tudták volna a megfigyelt mintát, ez a NOTMAR adathiány mechanizmus jellegzetessége.

ÚJ EREDMÉNYEK

I. tézis (publikáció: Máder 2005)

Az adathiányok különböző okoknál fogva keletkeznek. Ha ezek random módon keletkeztek, akkor jelentős hatást nem gyakorolnak a megfigyelhető esetek eloszlására. Ha ezek mögött viszont szisztematikus okok állnak, akkor az adathiány mechanizmusa a megfigyelhető eseteket befolyásolja. A következtetéseink torzzá válnak. A torzítatlanságtól a torzítottság felé három szintet különböztethetünk meg, ezek a MCAR, az MAR és a NOTMAR mechanizmusok. Az imputálások hatékonysága különböző a három adathiány mechanizmus mentén.

II. tézis (publikáció: Máder 2005)

Az egyre növekvő mértékű adathiány egyre növekvő adathiány kezelési igényeket támaszt. Minél nagyobb mértékű az adathiány és minél inkább szisztematikus az adathiány mögött álló ok-rendszer annál nagyobb a megfigyelhető esetek torzítottsága. A megfigyelt esetek adta becslések egyre torzulnak.

III. tézis (publikáció: Máder 2004)

Kiegészítettem az Oravecz Beatrix által definiált adathiány mintázatok 5 legjellemzőbb típusát egy, az adathiányos változók mögötti szociológiai magyarázat bevezetésével. Az MAR adathiány-mechanizmus esetében kimutattam, hogy bizonyos adathiányok szorosan együtt járhatnak, egységes adathiány mintázatot vesznek fel, és a mögöttük álló válaszadói magatartás és attitűd nevesíthető.

IV. tézis (publikáció: Máder 2005)

Az imputálási eljárások csökkentik a hiányos adatbázisok által adott becslések torzítottságát. Az MAR és a NOTMAR adathiány mechanizmusok esetében az adathiánnyal rendelkező adatbázisokból adott becslések torzak. Léteznek viszont olyan szofisztikált imputálási eljárások, melyek az adathiányos esetek átlagát, illetve heterogenitását jól imputálják. A különböző adathiány mechanizmusok esetében a HOT DECK eljárás bizonyult a leghatékonyabb imputálásnak, ez az eljárás a törölt esetek eloszlásának mind az átlagát, mind a szórását jól becsülte.

V. tézis (publikáció: Máder 2005)

Az imputálási módszereket Laaksonen négy fő kategóriába osztotta (Laaksonen, 1999). Megmutattam, hogy a model-donor imputálások és a donor imputálások a hatékonyak, a naiv megközelítések elégtelen eredményre vezetnek.

VI. tézis (A disszertációban az 5.2. fejezet.)

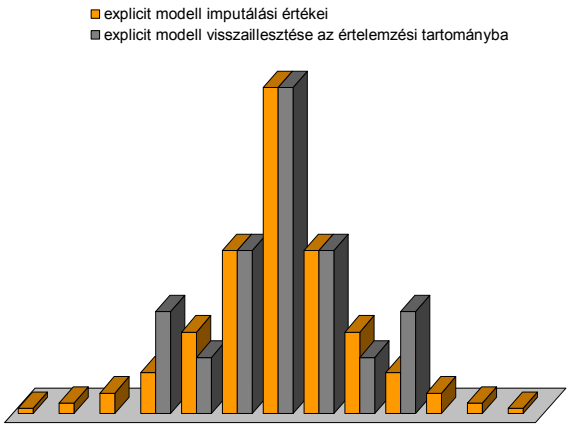
Az imputálási eljárásnak alapja lehet explicit függvény, implicit függvény vagy ezen függvények keveréke. Az implicit modellek olyan függvények, amelyek a megfigyelt, ismert, az adatbázisban megtalálható elemeket használják fel. Az explicit modellek pedig az implicit modellekhez hasonlóan az adatbázisban megtalálható elemeket, információkat használják fel a becslési algoritmushoz, de azokból csak a becslés modelljét készítik el. Az imputálni szándékozott értéket pedig ezen becslési algoritmus alapján extrapolálják. A közöttük lévő különbségek:

- magas mérési szintű és MCAR adathiány mechanizmusú imputálandó változó esetében a implicit modellek hatékonyabbak, mint az explicit eljárások
- alacsony mérési szintű és MAR adathiány mechanizmusú imputálandó változó esetében viszont az explicit modellek voltak hatékonyabbak, mert az adathiánnyal való összefüggések orientálták ezeket az eljárásokat

VII. tézis (A disszertációban a 8.5. fejezet.)

A REG regressziós becslésének természetes velejárója, hogy az eredeti tartományon túlmutathat, azaz az AC értelmezési tartományának területén kívüli extrém értékeket is adhat. Ennek az extrapolációnak a következménye, hogy az extrém értékek visszakódolódhatnak az AC eloszlás elvi minimumára illetve maximumára.

Az explicit modellek extrém értékeinek hatása



VIII. tézis (A disszertációban a 10.1. fejezet.)

A NOTMAR adathiány-mechanizmusának korrigálása alapvetően három lépcsőben történik, az első lépésben más, független adatfelvételek eredményeit kell begyűjteni, hogy információt kapjunk a torzulás mértékéről és jellegzetességeiről (1), a második lépésben egy deflációs függvényt kell előállítani a változó torzulásának kijavítására (2), majd végül ezen a korrigált adatbázison kell az adathiányos értékeket imputálni.

HIVATKOZÁSOK

Laaksonen, S. (1999) *How to Find the Best Imputation Technique?* Draft for the 1999 International Conference on Nonresponse, Portland, Oregon, November 28.

Little, R. J. A. – Schenker, N. (1995) Missing Data. In Arminger, G. – Clogg, C. – Sobel, M. (eds.) *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. New York: Plenum, p. 39–75.

Máder MP. (2005) Imputálási eljárások hatékonysága. *Statisztikai Szemle* 83 (7) p. 628–644.

Máder MP. (2004) Adathiány mintázatok az életeseményekre adott válaszok között. *Új Ifjúsági Szemle* 2 (4) p. 25–32.

McDermit, M. – Funk, R. – Dennis, M. (1999) *Data Cleaning And Replacement of Missing Values*. Kézirat.

Oravecz B. (2008) *A szelekciós torzítás és csökkentése az adósminősítési modelleknél*. PhD értekezés. Budapest: Corvinus Egyetem, Gazdálkodástudományi Doktori Iskola.

Rubin, D. B. (1976) Inference in Missing Data. In: *Biometrika* (63) p. 581–582.

Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley.

Rudas T. (2005). *Mixture Models of Missing Data.* Budapest: ELTE and TARKI. http://statisztika.tatk.elte.hu/tanszeki_honlap/RT_Q&Q.pdf.