ХУГАЦААН ЦУВААНЫ ГЭЭГДСЭН ӨГӨГДЛИЙГ НӨХӨХ НЬ: Улаанбаатар хотын агаарын чанарын цаг тутмын мэдээний өгөгдлийн сангийн жишээн дээр

Ч.Долгоржав
Монгол Улсын Боловсролын Их Сургууль,
Математик Байгалийн Ухааны Сургууль,
Мэдээлэл Зүйн тэнхим
dolgorjav@msue.edu.mn

Хураангуй
Цаг агаар, орчны агаарын бохирдлын ажиглалтын харуулууд автомат ажиллагаатай учир өгөгдөл цуглуулах шатанд техникийн хэвийн ажиллагаа доголдох, цахилгааны саатал г.м шалтгаанаар өгөгдлийн цуваа дутуу бичигдэх асуудал байнга тохиолддог. Орхигдсон өгөгдлийг нөхөж гүйцээх нь туршилтын үр дүнг сайжруулдаг ч ямар зарчмаар гээгдсэнийг нь тодорхойлж, оновчтой арга ашиглах нь чухал. Улаанбаатар хотын хувьд, агаарын чанарын мэдээ боловсруулахад анхан шатны өгөгдөлд боловсруулалт хийхдээ орхигдсон өгөгдлийг тооцдоггүй бөгөөд энэ ажлаар бид уг санд агаарын хэм (*цаг уурын*), хүхрийн хүчил, том ширхэглэлт тоосонцор (*агаарын бохирдлын*) үзүүлэлтүүдийн хоосон утгуудыг нөхөж, албан ёсны мэдээтэй харьцуулалт хийж үзлээ. Хэмжилтийн өгөгдлийн олонлогт *MI* буюу олон алхамт тооцооллын аргаар бодолт хийх нь илүү үр дүнтэй нь харагдлаа.

Түлхүүр үг: орхигдсон өгөгдөл, өгөгдөл алдагдах механизм, олон утгаар нөхөх

I. Оршил

Байгаль дээрх юмс үзэгдлийн нөлөө, цаг агаарын таагүй нөхцөл, техникийн доголдол, хүний санаатай болон санамсар болгоомжгүй үйлдэл зэрэг асар олон шалтгаанаар өгөгдөл орхигдох, алдаатай бичигдэх тохиолдлууд аль ч төрлийн судалгааны өгөгдлийн санд байнга таардаг. Ингэж дутуу, алдаатай бичигдсэн өгөгдөл нь туршилтын үр дүнг санаанд оромгүй өөрчилж болно. Байгаль орчин, цаг уурын өгөгдөл анхдагч хэлбэрээрээ алдаагүй байна гэж бараг үгүй бөгөөд хөгжиж буй буюу буурай хөгжилтэй орнуудын хувьд алдаатай өгөгдлийн бүртгэгдсэн хэмжээ өндөр, зарим тохиолдолд байх ёстой өгөгдлийн ердөө 30% нь бүртгэгдсэн тохиолдол байгаа нь ажиглагджээ [8], [9], [22], [25]. Ийнхүү алдааны хэмжээ ихсэх тусам бодит байдлыг ойлгох, шинжилгээ хийх болон таамаглал дэвшүүлэхэд хүндрэл учирна гэдгийг судлаачид олонтаа дурдсан байна [9], [24]. Үүнээс зайлсхийхийн тулд судалгаанд ашиглагдаж байгаа өгөгдлийг урьдчилан боловсруулдаг бөгөөд үүнд алдаатай, дутуу өгөгдлийг цэвэрлэх; өгөгдлүүдийг нэгтгэх; нормальчлах; өгөгдлийг багасгах; дискретчлэх үйлдлүүд багтдаг бөгөөд бид энэ удаад зөвхөн орхигдсон өгөгдлийг нөхөх аргын талаар авч үзлээ. Өгөгдлийн дутууг нөхөхдөө зөв арга сонгоогүйгээс болж үнэнд ойр биш утга бодогдох, энэ нь цаашлаад туршилтын үр дүнд сөргөөр нөлөөлнө. Иймээс гээгдсэн, орхигдсон өгөгдлийг нөхөхийн тулд тухайн алдагдал ямар төрлийнх болохыг нь эхлээд оношлох хэрэгтэй, ингэснээр түүндээ тохирох аргуудыг хэрэглэх боломж бүрдэнэ [7].

Өгөгдөл алдагдсан дүр зургаар 1) *univariate* буюу бүх хувьсагчид утгатай байхад зөвхөн нэг баганын утга алдагдсан байх; 2) *unit nonresponse* нь бүлэг хувьсагчид ижил тохиолдолд зэрэг утга бичигдээгүй, бусад нь бичигдсэн байх; 3) *monotone*. Хэмжигдэх хувьсагчид X_1, X_2, …, X_j; $j = 1, 2,\ldots, k\text{-}1$; $k > 1$ байг. Тэгвэл X_k, $k > j$ хувьсагч хэмжигдсэн бол X_j хэмжигдсэн байна; гэсэн хэлбэрүүдтэй байна. Энэ хэлбэрүүдийн аль нь тухайн тохиолдолд таарч байгаагаас хамаарч хоосон өгөгдлийг нөхөх аргаа сонгох нь зүйтэй.

Өгөгдөл дутуу бичигдэх 3 үндсэн шалтгааныг Rubin нар (1976) тодорхойлсон бөгөөд өнөөдөр ч хэрэглэгдсээр байна. Үүнд: 1) *missing completely at random* (MCAR). X хувьсагчийн орхигдсон утга X-ийн бусад утга болон, бусад хувьсагчдын утгаас, өөрөөр хэлбэл өгөгдөл дутуу бичигдсэн үзэгдэл нь ажиглагдсан утга болон ажиглагдаагүй утгаас аль алинаас хамааралгүй байх үед тохиолдох ба ийм төрлийн алдагдал ховор, боловсруулахад хялбар байдаг; 2) *missing at random* (MAR) - Алдаатай өгөгдөл, тухайн бичигдээгүй орхигдсон утгаас биш, харин бусад ажиглагдсан утгаас хамаарсан байвал энэ төрлийн алдагдалд тооцно; 3) *missing NOT at random* (MNAR) - Тодорхой шалтгаанаар алдаа гарсан байх, өөрөөр хэлбэл алдагдсан утга нь яагаад алдаа гарах болсонтой, тэр бичигдэх байсан дутуу орхигдсон өгөгдлөөс хамааралтай байх үед тохиолдоно. Энэ төрлийн алдагдал шалтгааныг тодруулах бэрхшээлээс хамаарч боловсруулахад хүндрэл их гардаг байна.

Бид энэ бүхнийг үндэслэж, өмнө дурдсан мэдээнд анализ хийж, орхигдсон өгөгдөл нөхөж гүйцээх аргуудыг туршсан талаар энэ өгүүлэлд авч үзнэ.

II. Арга зүй, материал

A. Судалгааны орчин, өгөгдөл

Улаанбаатар хотод Цаг уур орчны шинжилгээний газар (ЦУОШГ)-ын харьяанд 14 суурин харуул ажиллаж, цаг уурын үндсэн үзүүлэлтүүдээс гадна Монгол улсын агаарын чанарын стандартад бохирдуулагч хэмээн тодорхойлогдсон бодисуудыг цаг тутам хэмждэг. Үүнээс SO_2, NO_x үзүүлэлтийг бүгд; CO, PM_{10} үзүүлэлтүүдийг 6 харуул; O_3, $PM_{2.5}$ зэрэг элемэнтүүдийг алаг цоог хэмждэг байна. Улаанбаатарын хувьд хүлцэх хэмжээ[1]-нээс хамгийн

[1] *Хүний амьсгалын түвшний агаарт байх бохирдуулах хортой бодисын хүний биед шууд болон дам нөлөө үзүүлэхгүй, ажиллах*

их давдаг бохирдуулагч бодисуудыг онцлон авч үзвэл, харьцангуй урт хүйтний улирлаас шалтгаалах хатуу түлшний шаталт, температурын инверсээс шууд хамааралтайгаар, 10-р сараас дараа оны 4 сарыг дуусталх хугацаанд, хүхрийн хүчил (SO_2), том ширхэглэлт тоосонцор (PM_{10}), нарийн ширхэгт тоосонцор ($PM_{2.5}$) бодисууд байна (*Агаарын чанарын мэдээ*). Дурдсан хугацаанд, хүний амьсгалах агаарын нэг шоо метр дэх агууламж Монгол улсын стандарт (*зөвшөөрөгдөх дээд хэмжээ-ЗДХ буюу хүлцэх хэмжээ*)-аас дээд тал нь 10-25 дахин ихэсч, хүний эрүүл мэндэд онц аюултай түвшинд очдогийг Дэлхийн банкны тайланд тэмдэглэсэн төдийгүй, ДЭМБ-аас тодорхойлсон хөгжиж буй орнуудын ЗДХ-тэй харьцуулахад мөн л давсан үзүүлэлттэй[2] байна. Бид 5 станцын 2011-2014 оны хэмжилтийн өгөгдөлд шинжилгээ хийж үзэхэд нийтлэг бүртгэдэг 12 төрлийн өгөгдлөөс хамгийн багадаа 6.24%, дээд тал нь 26.15% дутуу байна (Хүснэгт 1). Энд, цөөн харуул дээр хэмжигддэг O_3 болон $PM_{2.5}$ бодисуудыг тооцоогүй болно.

утгууд маш өндөр, олон давтамжтай, дээд тал нь 9360 мкг/м^3 гэж бүртгэгдсэн байна. Дэлхийн банкны судалгаанд энэ талаар, Улаанбаатарын зарим харуул дээр PM_{10} хэт өндөр концентраци хэмжигдсэнийг тайлбарлаж чадаагүй, алдаа байх магадлалтай хэмээн дурджээ [2]. Эдгээр алдагдал болон алдаатай өгөгдлүүд бичигдэж байгаа шалтгаан нь цахилгааны саатал, харуул шинэчлэх, автомат багаж төхөөрөмжүүдийг суурилуулах, сорьц авах багаж эвдрэх, багажинд үзлэг үйлчилгээ хийх болон *шүүлтүүр бодисын солих хугацаа хэтэрсэн байх* (мэрг)[3] г.м өртөө харуулууд дээрх техникийн бүрэн бүтэн байдал, хэвийн ажиллагаанаас хамааралтай байна (Агаарын чанарын албаны тайлан). Манай улсад өнөөдрийн байдлаар нийслэлийн агаарын чанарыг мэдээлэх боловсруулалт хийхдээ дээрх алдааг тооцолгүйгээр, шууд станцууд дээрх хэмжигдсэн өгөгдлийг авч дундажлан тооцоолдог; ЦУОШГ нь цаг тутмын мэдээ мөн л нөхөлт хийдэггүй, урт хугацааны цуваан дээрх алдааг ойролцоо станцуудынх нь утгаар нөхдөг (мэрг) байна.

Хүснэгт 1. Станцууд дээрх мэдээний алдагдал. Өгөгдлийн хамрах хугацаа 2011-2014

Хэмжигдэх үзүүлэлт	Станц					Нийт алдагдал	Алдагдлын хувь
	UB2	UB4	UB5	UB7	UB8		
Агаарын даралт	2959	5305	1333	1674	10115	21386	**12.20**
Агаарын хэм	2959	3935	1047	1674	5321	14936	***8.52***
Хур тунадас	2227	3230	755	1252	4630	12094	***6.90***
Харьцангуй чийг	2959	5305	1047	1674	10115	21100	**12.04**
Салхины чиг	2973	5305	1047	1674	10115	21114	**12.04**
Салхины хурд	0	158	364	66	1537	2125	***1.21***
CO	4736	6221	7783	5842	11924	36506	**20.82**
NO_2	6711	8062	1833	4670	8741	30017	**17.12**
NO	6711	8062	1833	4670	13233	34509	**19.68**
NO_x	6711	6699	1833	4670	8741	28654	**16.34**
PM_{10}	4430	6998	4778	7888	12866	36960	**21.08**
SO_2	5270	6627	2605	4869	12706	32077	**18.30**
Нийт алдагдлын хувь	***11.10***	**15.04**	***5.99***	***9.27***	**25.12**	**13.31**	

Зөвхөн 2010 оны сүүлийн хагас дахь мэдээний өгөгдлийн алдагдал эдгээр харуулуудад илүү өндөр хувьтай, харгалзан 23.21; 75.28; 14.27; 15.47; 76.12 (%) байна. Аль ч нэгтгэлээс, 5 болон 7 дугаар харуулууд харьцангуй тогтвортой ажилладаг болох нь харагдаж байна.

Хүснэгт 2 зөвхөн огт тэмдэглэгдээгүй, хоосон орхигдсон өгөгдлүүдийг харуулж байгаа юм. Гэтэл байх ёстой дээд/доод хэмжээнээс хэтэрсэн өгөгдөл мөн л бүртгэгдэж байна. Авч үзэж байгаа өгөгдлийн санд агаар бохирдуулагчийн агууламж 455 удаа сөрөг тоо бичигдсэн байна. Энэ нь нийт өгөгдлийн тоотой харьцуулахад бага тоо боловч огт бичигдэх боломжгүй, худлаа утга орсон байгаа нь өдрийн, сарын болон жилийн дундаж мэдээ гаргахад шууд нөлөөлнө. Мөн, PM_{10} бодисын цагийн дундаж агууламж нь Монгол улсын агаарын чанарын стандартад 100 мкг/м^3 байхад дурдсан хугацааны өгөгдлийн сангаас харахад станц бүр дээрх хамгийн их

B. Өгөгдөл нөхөх аргууд

Алдаатай өгөгдлөөр туршилт хийхээс зайлсхийх олон арга замууд байдаг бөгөөд хамгийн энгийнх нь түүнийг огт тооцохгүйгээр хэвээр нь орхих явдал юм гэж судалгаануудад дурджээ [14]. Зарим судлаачид, алдаатай өгөгдлийг бүр хасах, эсвэл хэвээр нь орхих нь оновчгүй аргаар нөхөхөөс илүү дээр гэж үзсэн байна [5]. Манай орны нөхцөлд ч олон төрлийн судалгаанууд хоосон өгөгдлийг орхиж, хэмжигдсэн өгөгдлийг дундажлах замаар тооцоо хийдэг байна (мэрг; Дэлхийн банкны тайлан). MCAR болон MAR төрлийн алдагдал нь орхиж болдог (*ignorable*) төрлийнх учир түүэмэл хэрэглэгддэг *listwise, pairwise* процедураар устгаж болох юм. Гэсэн хэдий ч Little, Rubin нар (2002) *listwise* аргаар устгах нь үр дүнг бууруулдаг болохыг тодорхойлсон учраас аль болох зайлсхийх хэрэгтэй [18]. Мөн, Schafer (1999) зөвхөн MCAR өгөгдлийг л устгаж болохыг тэмдэглэсэн байна. Ажиглалтын утгууд нь хэмжигдэж байгаа хувьсагчдын хоорондын хамаарлыг илэрхийлж байдаг тул устгах арга бол тэр чухал мэдээллийн нэг хэсгийг давхар алдана гэсэн үг [15]. Ингэснээр нөхцөл байдлыг ойлгох, таамаглах г.м

чадвар, биеийн байдал, сэтгэхүйд харшлахгүй байхуйц агууламж, зөвшөөрөгдөх дээд хэмжээ. (Монгол улсын агаарын чанарын стандарт)

[2] *Монгол улсын агаарын чанарын стандарт PM_{10} хувьд 50 мкг/м^3,SO_2 хувьд 20 мкг/м^3, NO_2 хувьд 40 мкг/м^{3}байхад, ДЭМБ-аас гаргасан хөгжиж буй орнуудын агаарын чанарын стандартад PM_{10} –ын хувьд 70 мкг/м^3 байгаа билээ.*

[3] *Мэргэжилтнээс ярилцлагын аргаар авсан судалгаа.*

дараагийн алхмуудад хүндрэл учирдаг [5]. Ийм учраас өгөгдлийн алдааг засварлах, орхигдсон өгөгдлийг үнэнд хамгийн ойр байж болох утгаар нөхөх нь хугацаан цувааны шинжилгээнд маш чухал юм. Тухайн тохиолдлоос хамаарч судлаач орхигдсон өгөгдлийг нөхөх аргуудаас сонгох хэрэгтэй болно.

Нэг утгаар нөхөх. Single Imputation (SI): Зүй нь бол энэ арга MAR, MCAR аль ч төрлийн алдагдалд ашиглаж болдог бөгөөд тооцоолсон нэг л утгаар хоосон утгыг солино.

1. Энгийн арифметик дунджийн арга. Хамгийн түгээмэл, ашиглагддаг, хялбар арга бөгөөд хэд хэдэн төрлийн дунджийн арга байна. Гэхдээ дундажлах аргаас аль болох зайлсхийхийг Michael Friendly болон олон судлаачид санал болгосон байна. Нийтлэг ашиглагддаг зарим аргуудыг дурдъя.
   - Өмнөх болон дараагийн утгын дундаж (Top and Bottom mean/Mean-before-after). Цувааны хоосон утгыг хугацааны өмнөх болон дараагийн давтамжид ажиглагдсан утгуудын дунджаар нөхөх арга. Гэхдээ энэ нь маш цөөхөн утга орхигдсон үед л тохиромжтойг олон судлаач илрүүлсэн байна [4]. Хэрэв дараалан авах бичлэг олдоогүй бол энэ хүндрэлээс гарахын тулд өмнөх өдрийн, дараа өдрийн тухайн цаг дахь утгуудын дунджийг авч болно. Гэтэл манай өгөгдөл сонгосон станц бүрт цаг тутмын өгөгдөл үргэлжлэн орхигдсон хэмжээ нь маш урт (*том цоорхойтой)* байгаа нь энэ аргыг өмнөх болон дараа хоногийн ижил цаг дээрх утгаар нөхөх тохиолдолд ч ашиглахад учир дутагдалтай болох нь харагдаж байна (Хүснэгт 2).
   - Бүх бичлэгийн дундаж (Mean substitution). Тухайн хувьсагчийн орхигдсон байгаа хоосон утгыг түүний хувьд хэмжигдсэн байгаа бүх утгын дунджаар нөхөх арга. Гэхдээ ингэж нөхөлт хийснээс хоосон орхисон нь дээр гэдгийг судлаачид санал болгосон байна [5].
   - Цагийн дундаж (Hour mean method). Li нар (1999) энэ аргыг хэрэглэхдээ тухайн харуулын өгөгдлийн цуваанаас орхигдсон цагтай ижил хугацаан дахь, жилийн бүх утгуудыг дундажлан тооцсон. Эдгээр аргууд нь дулааны болон хүйтний улирлын агаарын хэмд цельсийн 70 орчим хэмийн зөрүү оршдог манай орны хувьд ялангуяа цаг уурын мэдээний хувьд ашиглахад тохиромжгүй юм.

Хүснэгт 2. Станцууд дээрх алдагдлын хамгийн урт цуваа

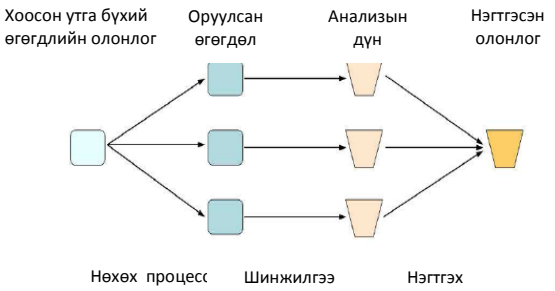
Станцын дугаар	UB2	UB4	UB5	UB7	UB8
Нийт алдагдлын тоо	48646	65907	26258	40623	110044
Дараалсан хамгийн урт алдагдал /хоногоор/	108.875	**249.79**	**202.08**	53.83	**214.91**

2. Ойр хөрш (ойр хөршүүдийн дундаж). Тухайн хувьсагчийн хувьд алдагдсан утгыг ойрын харуулын ижил хугацаан дахь утгаар шууд солих. Үнэнд ойр утга өгөх магадлалтай учир энэ арга нэлээн түгээмэл хэрэглэгддэг ч орчны хувьсагчид тогтвортой хэмжигддэг, хамааралтай утга өгч чадах ойрын хөрш байгаа үед л ашиглах нь зүйтэй [13]. Улаанбаатар хотын хувьд бичил уур амьсгалын мужууд, тэдгээрийн ялгаа аль хэдийн байршлаас хамааран бий болсон [3]; харуулууд өөр хоорондоо зайтай; өртөө харуулуудын байршил хот байгуулалтын болон газарзүйн ялгаатай орчинд байрладаг; ойролцоо харуулын ажиллагаа найдвартай биш зэргээс шалтгаалан төдийлөн оновчтой сонголт биш байж болох талтай гэж үзлээ. Манай өгөгдлийн санд дээр дурдсан агаар бохирдуулагч 6 элемэнтийн хувьд, хугацааны нэг давтамжид бүх станц дээр зэрэг алдагдах үзэгдэл нэлээд байна. ЦУОШГ-ын хувьд урт цуваанд энэ аргыг ашиглан нөхөлт хийдэг байна (мэрг).
3. Normal Ratio (NR) - Станцууд дээр ижил хугацаанд бүртгэгдэж байгаа утгуудыг харьцуулахад, зарим тохиолдолд (салхины хурд, хур тунадас г.м) тэдний пропорц, заримдаа тэдний ялгаа нь ч тогтмол байдаг (WMO[4] 1983). NR арга нь хэмжигдсэн утгын пропорцын (*жингийн*) тусламжтайгаар хоосон утгыг нөхдөг Өөрөөр хэлбэл тухайн хувьсагчийн хувьд хамгийн их давтагдаж хэмжигдсэн утгаар хоосон утгыг нөхнө.
4. Hot deck[5] (Case substitution) - Тухайн өгөгдлийн сан дотроос яг ижил нөхцөлд үүссэн өгөгдлийг хайж олон түүгээр нөхөх. Hox (1999) энгийн аргаар нөхөлт хийх аргуудаас хамгийн сайн үр дүн өгснийг дурдсан байхад Gold, Bentler нар (2000) хазайлтын коэффицент, стандарт алдаа нь өндөр байсныг дурдсан байна. Гэвч манай жишээнд, хугацааны интервал дахь байх ёстой өгөгдлийн ¾ алдагдсан тохиолдол ч байгаа учир мөн л хүндрэл үүсэх магадлалтай.
5. Интерполяц (шугаман, квадрат, куб, сплайн интерполяци). Norazian нар хэд хэдэн судалгаандаа (2002; 2006; 2008), Niska (2004), Junninen (2004) нар ч мөн эдгээр аргуудыг ашигласан бөгөөд эдгээрээс харахад өгөгдлийн алдагдал бага байх үед (нийт өгөгдлийн 15% хүрэхгүй хувь орхигдсон байхад) интерполяцийн арга сайн үр дүн өгсөн байна.

Олон утгаар нөхөх: Дээрх аргууд ердөө ганцхан ширхэг үнэлэгдсэн утгаар хоосон утгыг солино. Харин үүнээс өөрөөр, MAR төрлийн өгөгдлийн алдагдлыг олон утгаар нөхөх *Multiple Imputation (MI*) аргыг Рубин анх (1977) санал болгосон байна. Өгөгдлийн сан дахь цоорхойг олон утгаар нөхөх аргын *гол зарчим нь*, хэмжилтийн бусад өгөгдөл болон тухайн нөхцөлд үндэслэж, хоосон нүднүүд бүхий өгөгдлийн олонлогтой ижил хэмжээс бүхий боломжит $k > 1$ ($k \in [3; 10]$ *түгээмэл)* тооны олонлогийг хоосон утгуудыг өөр өөр аргаар тооцоолж дүүргэсэн байдалтайгаар үүсгэнэ. Дараа нь хамгийн магадлал өндөртэй эдгээр өгөгдлүүдэд дахин анализ хийж гаргасан утгаар эх олонлог дахь хоосон утгыг солино [27], [15], [7].

[4] *World Meteorological Organization*

[5] *Cold deck нь тохируулсан нэг утгаар нөхөх ба энэ утга нь өөр өгөгдлийн олонлогоос сонгогдсон байдаг учир төдийлөн ач холбогдолтой биш юм.*



Зураг 1. Олон утгаар тооцоолон нөхөх аргын ерөнхий диаграм

6. Multiple Imputation based on Markov Chain Monte Carlo (MCMC) Олон утгаар нөхөх гэдэг бол үндсэндээ Monte Carlo-гийн арга (Schafer, 1999) бөгөөд MCMC нь Марковын цувааны тархалтаас дуураймал тоо санамсаргүйгээр үүсгэдэг тоон арга юм [27]. Марковын цуваа X_0, X_1, X_2, … X_i хэлбэртэй дурын хувьсагчуудтай ба, элемэнт бүрийн тархалт хамгийн сүүлийн элемэнтээс хамаарна.

$$P[X_i < x \mid X_0 = x_0, X_1 = x_1, \dots, X_{i-1} = x_{i-1}] = P[X_i < x \mid X_{i-1} = x_{i-1}]$$

Эхлээд, хэмжигдсэн утгуудын дундаж вектор (μ), вариац-ковариацийн матриц (∑) ЕМ аргын адил бодогдоно. Өгөгдөл оруулах шатанд (*Imputation*), хэмжигдсэн өгөгдлүүдээр $Y_{i(obs)}$ өгсөн орхигдсон өгөгдлүүдийн $Y_{i(mis)}$ нөхцөлт тархалтаас дурын утга сонгож орхигдсон утгуудыг төсөөлнө. Дараагийн шатанд (*Posterior*), ковариацийн параметр болон дунджийн тархалтаар утгыг өөрчилж хадгална. Өөрчлөгдсөн параметрт тулгуурлан, хожуу тархалтаас ковариацийн матриц болон дундаж векторыг дахин бодно [27].

Schafer (1997) энэ аргыг 1) нормаль тархалттай; 2) MCAR буюу MAR төрлийн алдагдалтай; 3) орхигдсон өгөгдлийн дүр зураг нь *monotone*, эсвэл дурын байхад ашиглах боломжтой гэж үзсэн байна. Allison (2012) MAR төрлийн алдагдалд илүү тохирно гэж дүгнэжээ.

7. Өгөгдлийн алдагдал MAR байхад олон утгаар нөхөх өөр нэг арга бол *Maximum Likelihood (ML)* юм. Нэг утгаар нөхөх аргуудтай харьцуулахад хазайлтгүй тооцоолол хийдгээрээ давуу бөгөөд тэдгээр аргуудаас ихэнхдээ илүү үр дүн үзүүлдэг байна (Graham 2009). ML аргыг ижил өгөгдлийн олонлог дээр ашиглахад, өгөгдлийн хэмжээ их, нөхөлт хийх тоо том байхад, ML болон MI аргуудын үр дүн ижил байна (Collins нар. 2001). Allison (2012) -ы хийсэн туршилтад ML аргыг хэдэн ч удаа хэрэглэсэн ижил үр дүн өгсөн бөгөөд энэ бол тухайн аргын нэг давуу тал гэдгийг тэр дурджээ.

8. ЕМ бол ML аргатай тун төстэй нэг алгоритм бөгөөд энэ аргыг Dempster нар (1987) хөгжүүлсэн [11], [10]. Алгоритм нь *expectation* (E), *maximization* (M) гэсэн 2 алхамтай.

E – хоосон утгын хүлээгдэж байгаа утгыг олж (*хэмжигдсэн утгын нийлбэр, квадратуудын нийлбэр, вектор үржвэрүүдийн нийлбэрийн тусламжтайгаар*), хамгийн сайн үнэлэгдсэнийг нь түүний оронд бичнэ. Хамгийн сайн үнэлгээ гэдэг нь бусад хувьсагчдын хэмжигдсэн утгууд дээр тулгуурлан бодсон регрессийн тэгшитгэлийн коэффицентоор тодорхойлогдоно [10], [28] Хэмжигдсэн болон алдагдсан аль ч утга дээрх нийлбэрүүд рүү орж шинэчлэгдэнэ.

M – Нийлбэр, квадратуудын нийлбэр, вектор үржвэрүүдийн нийлбэр нь өмнөх алхамд нэгэнт үнэлэгдсэн учир ковариацийн матрицыг тодорхойлж, шаардлагатай регрессийн тэгшитгэлийг бодно. Дараагийн хоосон өгөгдөл орж ирэхэд, энэ регрессийн тэгшитгэлүүдийг ашиглан дараагийн хамгийн сайн үнэлгээг өөрчилнө. Нийлбэрүүдийг дахин тооцоолсоны дараа шинээр коварианцийн матриц болон дундаж векторыг бодно. Ингэснээр дараагийн алхамыг үнэлэх регрессийн тэгшитгэл үүснэ. Варианц, коварианц болон дундаж векторын өөрчлөлт маш бага болох үед үйлдэл зогсоно. Хэмжигдсэн, нөхөгдсөн өгөгдлүүдийн дундаж болон ковариацийн матрицаас шинэ утгаа гаргаж [10], [28].

ЕМ арга MAR төрлийн алдагдалд илүү тохирдог байна [7].

III. Туршилт, үр дүн

Олон судлаачид хугацаан цувааны алдагдсан утгуудыг нөхөхдөө тооцоололд суурилсан, олон утгаар нөхөх аргуудыг санал болгосон байна [15], [29], [17]. Гэхдээ туршилт хийхээс өмнө, эхлээд ямар шалтгаанаар өгөгдөл алдагдсан, ямар дүр зураг харагдаж байгааг мэдэх хэрэгтэй болно.

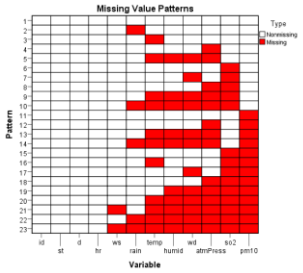
Өгөгдөл алдагдсан шалтгаан. Өмнө дурдсанаар, судалгааны ажиглалтын өгөгдөл хоосон байх шалтгаан нь MNAR байх нь элбэг хэдий ч манай жишээн дээр энэ төрлийнх байх боломжгүй юм. MNAR бол бичигдэх байсан утгаас шалтгаалж, түүнийг бичилгүй хоосон орхих нөхцөл байдаг. Тэгвэл, хэмжих өгөгдлийн дээд болон доод хязгаараас давсан буюу хэт хэлбэлзэлтэй өгөгдөл гарч ирэхэд утгыг бичих боломжгүй, хоосон орхидог гэж үзье. Гэтэл манай өгөгдлийн санд зөвшөөрөх утгын интервалаас гадна талд байрлах утга орсон тохиолдол олон байна (*сөрөг утга; дээд хэмжээнээс олон дахин давсан утга*). Иймд MNAR биш болж байна.

Aslan нар (2008) хур тунадасны бүртгэлийн мэдээний алдагдлыг MAR механизмаар үүсэлтэй гэж үзсэн байхад, Nurulkamal Masseran (2012) нарын судалгаа салхины хурд хувьсагчийн утга өгөгдлийн санд бичигдээгүй шалтгаан нь MCAR механизм гэж онцолсон байна. Манай нөхцөлд хүний хүчин зүйлээс, эсвэл цахилгааны саатлаас шалтгаалж харуул удаан хугацаагаар ажиллахгүй байх тохиолдол элбэг (мэрг). Бидний туршилт хийхээр авсан өгөгдлийг бүхэлд нь Little's MCAR Test (Roderick J. A. Little)-ээр шалгаж үзэхэд MCAR биш болох нь батлагдлаа (*p-value<0.05*).

Судлаачид байгаль орчны болон агаарын бохирдлын өгөгдлийг цуглуулах станцууд дээрх өгөгдлийн алдагдал ерөнхийдөө MAR зарчмаар алдагддаг гэж дурдсан байна [10], [4]. Schafer, Graham нар (2002): *"Хэрэв өгөгдөл алдагдах нь судлаачийн хяналтаас гадуур, тархалт нь тодорхойгүй л бол байж болох ганц хувилбар бол MAR"* гэжээ.

Олон утгаар нөхөх арга ашиглан хугацаан цувааг нөхөх нь нэг утгаар нөхөх аргуудаас нь алдаа бага гардаг болохыг олон судалгаа баталсан байна. MAR болон MCAR механизмаар гээгдсэн өгөгдлийн хувьд ML болон MI аргуудыг олон судлаачид санал болгожээ [6], [5], [17].

Алдагдлын дүр зураг. Бидэнд байгаа өгөгдлийг алдагдлын дүр зургаар нь шинжиж үзэхэд, холимог хэлбэртэй байгаа бөгөөд (Зураг 1), ийм нөхцөлд олон утгаар нөхөх аргыг ашиглах нь илүү үр дүнтэй болох нь харагдсан байна [10]. *Monotone* хэлбэр дээр Monte Carlo Markov Chain (MCMC) арга илүү тохирч буй нь харагдаж байна [19], [18].



Зураг 2. Дурдсан өгөгдлийн сангийн хоосон утгууд.

Туршилт. Бид EM болон MI аргуудыг ашиглан туршилт хийхээр сонгон авлаа. Туршилтын эхний алхамд, эдгээр аргуудаас тохирох аргыг сонгохын тулд 10 болон 20 хувийн алдагдал бүхий өгөгдлийн санг зохиомлоор үүсгэж, нөхөлт хийж, үр дүнд харьцуулалт хийв. Ингэхдээ өгөгдлүүдийг холимог хэлбэртэй, өөрөөр хэлбэл мөрийн болон баганын дагуу аль алинд нь дараалсан алдагдал байхаар тооцож хасалт хийв.

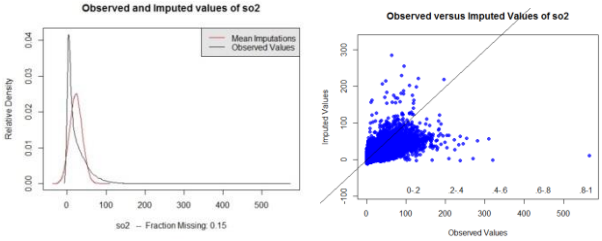
Дараа нь үнэнд хамгийн ойр утга гаргаж байгаа аргыг сонгон ЦУОШГ-ын өгөгдлийн санд байгаа хэмжилтийн утгуудыг нөхсөний дараа, дундаж утгыг тооцоолон өдрийн болон сарын дунджийг тэдний өөрсдийнх нь гаргасан тооцоотой харьцууллаа. ЦУОШГ-аас авсан 5 станцын 4 жилийн мэдээн дээр боловсруулалт хийн, жилийн аль өдөр, өдрийн аль цагт байгаагаас хамаарч өөрчлөгдөж байгаа хамаарлыг тусгахын тулд хугацаан цуваа руу дурдсан хоёр хувьсагчийг нэмж оруулсан. Гэхдээ статистик тайлбараас харахад SO_2 бохирдуулагчийн агууламжийг илэрхийлэх утга эдгээр хувьсагчдаас хүчтэй хамаарахгүй байна. Туршилтын өгөгдлийг MySQL Server дээр нэгтгэн, боловсруулалт хийгээд, R программчлалын хэлний (*mi, mice, amelia packages*) тусламжтайгаар нөхөлтийг гүйцэтгэлээ. Анализыг R болон SPSS программуудын тусламжтайгаар гүйцэтгэсэн болно.

Үр дүн. Цаг агаарын болон орчны агаарын бохирдлын мэдээллийн хугацаан цувааны алдагдсан өгөгдөлд нөхөлт хийх аргуудыг туршиж үзэхэд MI (*multiple imputation*) бага алдаатай ажиллаж байна (Хүснэгт 3).

Хүснэгт 3. EM, MI аргуудыг 10% өгөгдлийг зохиомлоор гээсэн олонлог дээр туршсан байдал

Аргууд	Үзүүлэлт	PM_{10}	SO_2	temp
Эх олонлог	Mean	794.25177	113.23759	-21.847872
Expectation Maximaztion	Mean	769.6028	113.7057	-21.9681
	RMSE	328.8887	10.34305	0.807259
	MAE	58.28723	1.716312	0.417376
	IA[6]	0.717331	0.747896	0.742282
Multiple imputation	Mean	783.8185	113.8595	-22.0218
	RMSE	328.8887	17.95669	2.589915
	MAE	58.28723	3.120222	0.749149
	IA	0.745261	0.74953	0.686951

Үүний дараа дурдсан өгөгдлийн санд байгаагаар хамгийн их алдагдалтай хэмжилт хийсэн UB8 станцын бодит өгөгдөл дээр *MI* аргаар туршилт хийж үзэхэд хэмжилтийн алдаа их байх үед нөхөлт төдийлөн хангалтгүй байгааг харуулж байна (Зураг 3а, 3б).



Зураг 3а, 3б. Хэмжигдсэн болон хоосон утгууд

Дараагийн шатанд 4 жилийн бүх цаг тутмын өгөгдлийг *MI* аргаар тооцохдоо доор дурдсан өгөгдлийг ашигласан болно (Хүснэгт 4).

Хүснэгт 4. Ажиглалтын 5 харуул дээрх мэдээ [7]

Variables	N	Mean	Std. Deviation	Missing	
				Count	Percent
atmPress	153809	866.56	7.371	21386	12.2
rain	163101	.03	.403	12094	6.9
humid	154095	56.67	18.576	21100	12.0
temp	160259	-.53	15.431	14936	8.5
wd	154081	164.66	112.711	21114	12.1
ws	173070	1.94	1.918	2125	1.2
pm10	138259	232.79	444.104	36936	21.1
so2	143143	26.63	44.035	32052	18.3

Бодолтын үр дүнгээс гарсан өгөгдлөөс өдрийн дунджийг 2014 оны байдлаар 5 харуулын хувьд тусад нь бодоод ЦУОШГ-аас ирсэн 2014 оны нэгтгэлийн харуул тус бүрийн дүнтэй харьцууллаа (Хүснэгт 5).

Хүснэгт 5. MI аргаар бодолт хийсний дараа SO_2 өдрийн дундаж мэдээг тооцож ЦУОШГ-ын мэдээтэй харьцуулав.

ST	Тооцоо	Valid	Missing	Mean	MAE	Median	Std. Dev
UB2	ЦУОШГ	294	71	22.908	1.0855	17.5	18.612
	MI	365	0	22.441	0.8903	18.125	17.01
UB4	ЦУОШГ	338	27	23.367	1.1817	15	21.726
	MI	365	0	23.515	1.1152	14.833	21.306
UB5	ЦУОШГ	327	38	29.061	1.7314	15	31.308
	MI	365	0	28.648	1.5541	18.583	29.69
UB7	ЦУОШГ	262	103	16.382	0.8063	14	13.051
	MI	365	0	18.41	0.9819	9.9167	18.76
UB8	ЦУОШГ	227	138	20.269			
	MI	365	0	18.737	0.7286	17.75	13.92

[6] *Index of Agreeent*

[7] *Бүх утга бүртгэгдсэн хувьсагчид, зарим харуул дээр бүртгэдэггүй үзүүлэлтүүдийг оруулаагүй учир анхны бүх өгөгдлөөр тооцсон хэмжээнээс зөрүүтэй гарлаа.*

IV. ДҮГНЭЛТ, ЦААШДЫН СУДАЛГАА

Монгол улс гадна орчны агаарын бохирдлын хэмжээгээр дэлхийд тэргүүлж байгаа өнөө цагт үнэн бодитой мэдээллээр иргэдийг хангах нь чухал асуудал юм. Үүнд хүрэх эхний алхам бол анхан шатанд бодит өгөгдөл цуглуулж байгаа нэгжийн үйл ажиллагааг хэвийн байдлаар хангах, түүнчлэн мэдээ боловсруулах түвшинд аль болох алдаагүй, цэвэр өгөгдөл ашиглан боловсруулалт хийх нь хамгийн зөв дүгнэлт гаргана.

Гэсэн хэдий ч өгөгдлийн алдагдал байх нь зайлшгүй учир энэ судалгаагаар хэмжилтийн алдааны нэг төрөл болох орхигдсон өгөгдлийг үнэнд ойр байж болох утгаар хэрхэн нөхөх талаар авч үзсэн бөгөөд олон утгаар харьцуулж нөхөх нь хамгийн оновчтой сонголт гэдэг нь харагдлаа.

Нөгөө талаар хангалттай урт цуваа байхгүй; өртөө харуулуудын хэмжилтийн мэдээ олдоц муутай; хэмжигдсэн өгөгдлийн чанар, үнэний баталгаа найдваргүй зэрэг шалтгааны улмаас судалгааны үр дүн гарцаагүй үнэн байх боломжгүй асуудлуудтай тулгарч байна.

Цаашид энэ ажлыг чанаржуулахын зэрэгцээ өгөгдлийг урьдчилан боловсруулах бусад аргуудыг туршсаны дараа бодит үр дүнд гарах өөрчлөлтийг авч үзэх болно.

ТАЛАРХАЛ

Энэ ажилд шаардлагатай мэдээллээр хангаж өгсөн Цаг уур орчны шинжилгээний газрын Орчны хяналт, шинжилгээний хэлтэст талархал илэрхийлье.

АШИГЛАСАН МАТЕРИАЛ

[1] МОНГОЛ УЛСЫН АГААРЫН ЧАНАРЫН СТАНДАРТ MNS 4585: 2007

[2] Дэлхийн банк. (2011). Улаанбаатарын агаарын чанарын дүн шинжилгээ. “Эрүүл мэндэд үзүүлэх сөрөг нөлөөг бууруулахын тулд агаарын чанарыг сайжруулах нь” судалгааны тайлан

[3] Оюуноров.Ж, (2011). Нийслэл Улаанбаатар хотын бичил уур амьсгал, түүний хандлага. УЦУХ-ийн Эрдэм шинжилгээний бүтээл. Уб., 2011 он

[4] Norazian Mohamed Noor, Yahaya Ahmad Shukri, Ramli Nor Azam, Abdullah Mohd Mustaf Al Bakri. (2008). Estimation of missing values in air pollution data using single imputation techniques. *ScienceAsia 34:* 341-345.

[5] Graham.J.W. (2012). Missing Data: Analysis and Design, Statistics for Social and Behavioral Sciences

[6] Newman.D.A. (2009). Missing data techniques and low response rates

[7] Joop J.Hox. (1999). A review of current software for handling missing data.

[8] Campozano.L., Sanchez.E., Aviles.A., Samaniego.E. (2014). Evaluation of infilling methods for time series of daily precipitation and temperature: The case of the Ecuadorian Andes

[9] Mohammed MRAOUA, Driss BARI. (2006). Temperature stochastic modeling and weather derivatives pricing: empirical study with Moroccan data

[10] Heikki Junninen, Harri Niska, Kari Tuppurainen, Juhani Ruuskanen, Mikko Kolehmainen. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment 38 (2004)* 2895–2907

[11] Nuradhiathy Abd Razak, Yong Zulina Zubairi, Rossita M. Yunus. (2014). Imputing Missing Values in Modelling the PM10 Concentrations (Mengganti Nilai Hilang dalam Pemodelan Kepekatan PM10)

[12] Aslan.S, Yozgatlig.C, Iyigäun.C, Batmaz.I, Täurkes.M, Tatli.H. (2008). Comparison of missing value imputation methods for Turkish monthly total precipitation data.

[13] Ceylan Yozgatligil, Sipan Aslan, Cem Iyigun, Inci Batmaz. (2012). Comparison of missing value imputation methods in time series: the case of Turkish meteorological data.

[14] Ahmad Shukri Yahaya, Nor Azam Ramli, Fauziah Ahmad, Norlida Mohd, Nor Muhammad, Nor Hakim Bahrim. (2011). Determination of the Best Imputation Technique for Estimating Missing Values when Fitting the Weibull Distribution. *International Journal of Applied Science and Technology*; November 2011

[15] James Honaker, Gary King. (2010). What to Do about Missing Values in Time-Series Cross-Section Data. *American Journal of Political Science, Vol. 54, No. 2, April 2010,* Pp. 561–581

[16] Atakan Kurt, Betul Gulbagci, Ferhat Karaca, Omar Alagha. (2008). An online air pollution forecasting system using neural networks. *Environment International* 34(2008) 592-598

[17] Paul D. Allison. (2012). Modern Methods for Missing Data.

[18] Michael Friendly. (). Missing data.

[19] Norazian Ramli M.N., Yahaya, A.S., Ramli, N.A., Yusof, N.F.F.M., Abdullah, M.M.A. (2013). Roles of Imputation Methods for Filling the Missing Values: A Review. *Advances in Environmental Biology, 7(12) October 2013*

[20] Joseph L Schafer. (1999). Multiple Imputation: a primer. *Statistical Methods in Medical Research. 1999;* 8; 3-15.

[21] J.L. Schafer. (1997). Analysis of Incomplete Multivariate Data. London: Chapman and Hall, Inc.

[22] Alvaro Escribano, Jorge Pena. (2009). Empirical Econometric Evaluation of Alternative Methods of Dealing with Missing Values in Investment Climate Surveys.

[23] Noor, N.M., Tan, C.Y., Abdullah, M.M.A., Ramli, N.A. & Yahaya, A.S. (2011). Modelling of PM10 concentration in industrialized area in Malaysia: A case study in Nilai. *2011 International Conference on Environment and Industrial Innovation IPCBEE,* Vol.12. Singapore: IACSIT Press.

[24] Noor, N.M. & Zainudin, M.L. (2008). A review: Missing values in environmental data sets. In *Proceeding of International Conference on Environment*.

[25] Noor, N.M., Yahaya, A.S., Ramli, N.A. & Abdullah, M.M.A. (2006). The replacement of missing values of continuous air pollution monitoring data using mean top bottom imputation technique. *Journal of Engineering Research & Education* 3: 96-105.

[26] Ryan W. Allen, Enkhjargal Gombojav, Baldorj Barkhasragchaa, Tsogtbaatar Byambaa, Oyuntogos Lkhasuren, Ofer Amram, Tim K. Takaro, Craig R. Janes. (2011). An assessment of air pollution and its attributable mortality in Ulaanbaatar, Mongolia. *Air Qual Atmos Health*.

[27] Lily Ingsrisawang and Duangporn Potawee. (2012). Multiple Imputation for Missing Data in Repeated Measurements Using MCMC and Copulas.

[28] Mark Huisman. (2010). Missing data

[29] Graham, J.W. (2012). Analysis of Missing Data