**MODELOS DE INFERENCIA NO PARAMÉTRICA PARA INDICADORES MEDIOAMBIENTALES**

1. Documento 1B-1: Resumo do proxecto
2. Documento 1B-2: Estado da arte da investigación proposta

Referencias Datos Direccionales:

Fisher,N.I., Lewis,T., and Embleton, B.J.J. (1993), Stadistical Analysis of Spherical Data, Cambrige: University Press.

Halll, P.; Watson, G. and Cabrera, J. (1987). Kernel density estimation with spherical data. Biometrika. **74**, 751-762.

Mardia,K., and Jupp, P. (2000), Directional Data, New York: Wiley

Zhao, L. and Wu, C. (2001). Central limit theorem for integrated squared error of kernel estimators of spherical density. Science in China (Serie A). 44, 474-483

**Recentemente xurdiron técnicas que poden ser de interese a hora de incluír variables meteorolóxicas como a velocidade e dirección do vento como os datos direccionais. Sen dúbida algunha, o libro que máis contribuíu a popularizar as técnicas estatísticas para datos direccionais é (Mardia, K., et al., 2000). Hall construíu o estimador núcleo da densidade con datos esféricos (Hall, P., et al., 1987) e máis recentemente (Zhao, L., et al., 2001) aplicouse o teorema central do límite para el error cuadrático medio estimador núcleo da densidade con datos esféricos; como ya apuntaba Hall en 1984.**

A preocupación pola degradación do medio ambiente e a actual lexislación sobre os niveis de polución, fan imprescindible a utilización de metodoloxía estatística para a predición de distintas sustancias no medio. En particular, para a predición de SO2 no entorno da Central Térmica de As Pontes utilizáronse modelos semiparamétricos (García Jurado, et al.,1995), modelos parcialmente lineales (Prada Sánchez,et al., 2000), modelos basados en redes neuronais (Fernández de Castro,et al., 2003), modelos con resposta binaria (Roca Pardiñas, et al., 2004), modelos aditivos (Cotos Yáñez, et al., 2004), modelos funcionais (Fernández Castro, et al., 2005).

No período 2008-2009 incorporouse unha nova Central de Ciclo Combinado o que causou un importante impacto nos sistemas de predición, xa que isto obrigou a cos esforzos metodolóxicos estivesen centrados no desenrolo, implementación e validación de novos modelos capaces de predicir non so niveis de SO2 sino tamén de NOX. Actualmente, están implementados os modelos aditivos para predicir con media hora de antelación o comportamento das variables de inmisión de SO2 y NOX. Estas predicións, se realizan de forma independente para cada un deles.

Ademais, a posta en marcha desta nova Central de Ciclo Combinado durante el período 2008-2009 provocou o deseño de novas metodoloxías de clasificación, as cales proporcionan con frecuencia minutal, a probabilidad estimada de co orixe dun episodio de alteración de calidade de aire débase á central térmica ou á de ciclo combinado. Á vista distes resultados, plantease o estudio con modelos de datos direccionales xa que podería aportar interesantes aplicacións na determinación do orixe dun episodio, como introducir variables meteorolóxicas no modelo; velocidade e dirección do vento.

1. Documento 1B-3: Obxectivos do proxecto

Neste proxecto existen tres vía de investigación principais. A continuación detállanse os obxectivos concretos a acadar en cada unha delas:

1. Elaboración de meteoroloxía dentro do campo dos datos funcionais.
   1. Estudio do modelo de regresión con covariables funcionais e resposta real o funcional.
      1. Modelo lineal xeneralizado (GLM)
      2. Modelo parcialmente lineal (LMP)
      3. Estudio aditivo xeneralizado (GAM)
   2. Estudio da robustez nos modelos 1.1.1., 1.1.2. y 1.1.3. robustos e con datos faltantes.
   3. Elaboración de contrastes nos modelos anteriores.
   4. Elaboración de test para a selección de modelos
   5. Extensión aos modelos con:
      1. Datos dependentes.
      2. Datos espacio-temporais.
      3. Datos direccionais.
   6. Estudio de simulación dos modelos anteriores.
   7. Aplicación dos modelos 1.1.1., 1.1.2. y 1.1.3. á predición de series temporais e espacio-temporais medioambientais con datos direccionais.
2. Aditivos
3. Elaboración de meteoroloxía dentro do campo de datos direccionais.
   1. Estudio de distribucións notables de datos direccionais.
   2. Estudio de modelos de predición con datos direccionais.
   3. Aplicación do estudado en 3.1. e 3.2. á predición de series medioambientais.
4. Elaboración de software
   1. Transferencia tecnolóxica da na metodoloxía á Central Térmica de As Pontes.
   2. Creación e optimización do novo software científico creado a partir do estudio desenrolado no apartado.

Anéxase un diagrama indicativo das posibles relacións entre as distintas liña de traballo.

1. Documento 1B-4: Aplicabilidade e posible transferencia dos resultados

Da frutífera relación entre a Sección de Medio Ambiente da U.P.T. As Pontes e o Departamento de Estatística e Investigación Operativa da Universidade de Santiago de Compostela nace o Sistema de Predición Estatística de Inmisión (SIPEI) debido al compromiso de Endesa co medio ambiente e o seu interese por mellorar este sistema de predición.

O Sistema de Predición de Estatística de Inmisión utiliza datos de inmisión minutales dispoñibles no sistema MEDAS para obter predicións dos valores de dióxido de xofre y de óxidos de nitróxeno, con media hora de antelación, usando modelos aditivos. Este sistema de predición tamén proporciónanos una predición probabilística de superación de los niveles de SO2 y NOx; tendo en conta os límites contemplados na lexislación vixente.

Ademais, SIPEI calcula a probabilidade estimada de se é a Central Térmica ou o Ciclo Combinado o causante do episodio de alteración da calidade do aire.

Nos últimos anos, o Sistema de Predición Estatística de inmisión adoptou unha estrutura modular na que se poden distinguir tres grandes bloques: a conexión, a predición e a visualización.

A conexión engloba todo o que ten que ver co orixe, a estrutura e a manera de obter os datos; constitue a unión entre o provedor dos datos e o propio sistema de predición. Neste módulo obtense os datos orixinais, formatéanse convenientemente y póñense a disposición del módulo de predición.

A predición fai referencia aos aspectos máis metodolóxicos, matemáticos y estatísticos; aquí encóntranse os procesos xurdidos da implementación dos modelos desenrolados. Neste módulo lense os datos formateados, aplícanse os modelos, realízanse os cálculos, obtense os resultados que se poñen a disposición do módulo de visualización.

A visualización comporta aspectos relacionados coa disposición e presentación da información. Neste módulo lense os resultados, xenáranse os gráficos e os informes e ponse toda a información a disposición do usuario final.

De todo isto conclúese co sistema é unha importante ferramenta de información e axuda na toma de decisións.

Anéxase unha imaxe coa pantalla principal da visualización.

1. Documento 1B-5: Plan de difusión e explotación de resultados

Os avances logrados durante a execución deste proxecto estarán ao alcance da comunidade investigadora e da sociedade en xeral, a través da publicación de Technical Reports (accesibles desde a páxina web do departamento http://eio.usc.es/pub/reports.html), de comunicacións en congresos e das tese doutorais dos bolseiros integrantes do equipo.

Tamén se publicarán artigos en revistas internacionais incluídas nos Journal Citation Reports tanto nas de carácter máis metodolóxico (Scandinavian Journal of Statistics, Canadian Journal of Statistics, Statistica Sinica…) como nas máis aplicadas (Technometrics, Environmetrics, Econometrica, Journal of Chemometrics, Computational Statistics and Data Analysis…). Por outra banda, dous membros do equipo pertencen a ERCIM (European Research Consortium for Informatics and Mathematics, http://www.ercim.org/) como coordinadores, xunto con Frederic Ferraty e Philippe Vieu da Universite Paul Sabatier de Toulouse, do tópico Statistics for Functional Data dentro do grupo de traballo Computing & Statistics (http://www.dcs.bbk.ac.uk/ercim/TrackSFD.html). Ademais o departamento pertence á rede europea IAP (http://www.stat.ucl.ac.be/IAP/PhaseVI/index2.html) que organiza workshops, seminarios, cursos,... Todo isto favorecerá a difusión internacional dos resultados obtidos e do software creado.

Para a construción do software empregarase a linguaxe de programación R (software libre) , ferramenta empleada para a creación dun paquete de datos funcionais que estará a disposisión de futuros usuarios no repositorio de contribuciones (http://cran.proyect.org). Ademáis aproveitando que o departamento pertence ao programa CONSOLIDER, incluirase no repositorio de software que se ten previsto crear dentro dito programa.

Todos os resultados que se van obter permitirán o estudo de datos medioambientais ou de tipo ecolóxico, a predición da evolución de indicadores medioambientais,... o que favorecerá a transferencia tecnolóxica de todos estes avances teóricos ao mundo empresarial (Posibles beneficiarios: ENDESA).

1. Documento 1B-6: Interese para o sector e posibles beneficios

Tanto a temática dos datos funcionais como a temática dos datos direccionais son de recente aparición polo que hai un amplo abanico de liñas de investigación aínda sen explorar.

Por iso un dos obxectivos deste proxecto será estender os modelos de predición xa construídos no ámbito unidimensional e multidimensional ao terreo funcional, é de grande interese desde o punto de vista teórico. Ademais esta metodoloxía pode proporcionar ferramentas para o tratamento de grandes cantidades de datos de carácter funcional de diversas disciplinas (Economía, Medio Ambiente, Medicina, Química,...), que os modelos existentes non son capaces de tratar de maneira adecuada.

Por outra banda, este proxecto tamén ten como obxectivo levar os datos direccionais o campo das series de tempo, o que é de gran interese tanto dende un punto de vista teórico como práctico. Ademais, isto sería beneficioso á hora de construír modelos de predición xa que podería incluír variables meteorolóxicas (por exemplo, dirección e velocidade do vento) ó longo do tempo para perfeccionar sistemas de predición ambientais.

1. Documento 1B-7: Metodoloxía

No que ao desenvolvemento matemático se refire, a metodoloxía a seguir será a habitual nestes casos empregando tanto ferramentas estatísticas como probabilísticas: deseño dun modelo con datos direccionais en series de tempo e estudo do seu comportamento con datos reais de control de calidade do aire, construción dun novo estimador para a regresión funcional e estudo do seu comportamento mostral e da súa distribución asintótica; elaboración de tests baseados en dito estimador e validación do seu comportamento mediante simulación; extensión das técnicas de aprendizaxe aos datos funcionais…

En canto ao aspecto computacional, o software que se constrúa estará escrito na linguaxe de programación R (software libre). O entorno de programación estatística R consolidouse como unha plataforma de desenrolo de algoritmos estatísticos avanzados. O entorno é libre e gratuíto o que o fai accesible á comunidade científica, académica e profesional os resultados das aplicacións e métodos implementados. As técnicas e tratamentos básicos xa están implementados no repositorio de contribuciones (http://cran.proyect.org).

Ademais, ao traballar con grandes conxuntos de datos, sobre os que se farán continuas remostraxes, xerarase unha carga computacional considerable que fará imprescindible o emprego de equipos informáticos veloces e con gran capacidade.

No relativo á metodoloxía de traballo propiamente dita, os sete membros do grupo reuniranse semanalmente. Por outra banda, as viaxes que unha das bolseiras realiza dúas veces á semana a Central Térmica de As Pontes, permitirán a transferencia tecnolóxica dos traballos realizados á área medioambiental, e a estreita relación entre o Departamento de Estatística e IO da USC.

Cómpre salientar que a relación entre o grupo e a Sección de Medio Ambiente da Central Térmica de As Pontes (a través de varios proxectos entre ENDESA e a USC, como se pode observar nos currículums dos membros do equipo) xa propiciou a transferencia tecnolóxica ao longo dos últimos anos (anéxase unha carta do xefe da Sección de Medio Ambiente da Central Térmica). Mostra disto é a aplicación de numerosas metodoloxías estatísticas para a predición de niveis de SO2 como se amosa no diagrama que se anexa.

1. Documento 1B-8: Plan de traballo
2. Documento 1B-11: Explicación e xustificación dos gastos
3. Documento 1B-12: Experiencia do equipo de investigación
4. Documento 1B-13: Instalacións e instrumentos

O departamento de Estatística e Investigación Operativa da USC, do que forman parte os membros do equipo, posúe un grid de computación paralela obtido na recente convocatoria de infraestructuras da Xunta de Galicia que consta de 11 servidores AMD opteron 2.4 GHz con dobre procesador dobre núcleo. Dito grid pode satisfacer as necesidades de simulación froito da execución deste proxecto.

Ademais cada un dos membros dispón dun ordenador persoal (cunha antigüidade media de dous anos).

Por outra banda, a bolseira do grupo que visita semanalmente a Central Térmica de As Pontes emprega alí un equipo informático pertencente ao departamento co que xestiona os datos medioambientais proporcionados pola Central (cunha antigüidade de 5 anos).

1. Documento 1B-14: outros medios

Debido as necesidades de velocidade e alta capacidade computacionais que precisa o correcto desenvolvemento do proxecto, téñense que renovar os cinco equipos informáticos dos membros grupo.

Ademais, sería preciso dispoñer dunha matriz de discos duros común para almacenar as inxentes cantidades de datos que se van a manexar e optimizar a utilización do grid do que xa dispón o departamento.

Polos mesmos motivos expostos anteriormente, tamén se precisa renovar o equipo informático que se ten nas instalacións da Central Térmica de As Pontes.