**MODELOS DE INFERENCIA NON PARAMÉTRICA PARA INDICADORES MEDIOAMBIENTAIS**

1. **Documento 1B-1: Resumo do proxecto**

A estatística non paramétrica e semiparamétrica, a análise de datos direccionais e a modelización con datos funcionais constitúen campos de máximo interese no entorno da Estatística na investigación actual a nivel internacional. No presente proxecto de investigación preténdese realizar novos desenrolos metodolóxicos integrados nestes tres campos e, con especial énfases, na aplicación e na transferencia ó contexto do Medio Ambiente, contribuíndo a determinar estratexias para preservar o entorno en base á análise de datos. Moitos dos novos métodos que se desenrolen serán implementados por primeira vez (empregando o software R) ou representarán xeneralizacións de versións anteriores (pasarase da predición de indicadores unidimensionais a multidimensionais). Ditos métodos serán validados mediante resultados teóricos e de simulación e, contrastados con mostras de datos reais que o grupo manexa grazas a súa relación permanente coa central térmica de As pontes (propiedade de Endesa Generación S.A) desde fai 22 anos. Tamén se deseñará unha ferramenta de clasificación que permita distinguir a orixe dun episodio de alteración de calidade e aire entre os posibles focos de contaminación no entorno da central, para o que será clave o tratamento con datos funcionais. Todas as técnicas propostas presentan un grande atractivo tanto desde o punto de vista académico, xa que a investigación no eido dos datos funcionais e direccionais son áreas de investigación emerxente, coma desde o punto de vista empírico, posto que na actualidade en diferentes campos das ciencias aplicadas (enxeñería medioambiental, química, biometría, medicina,...) se recolectan grandes cantidades de datos o que fai necesario a formulación multidimensional dos problemas ou a consideración de que estes se poidan aproximar por curvas. Os métodos obtidos ó longo deste proxecto poderán ser aplicados no contexto do Medio Ambiente grazas a ampla colaboración entre o noso grupo de investigación e a central térmica de As Pontes, como xa se comentou, e serán de potencial transferencia a outros contextos: produción con enerxías renovables (especialmente a eólica, que usa metodoloxía de datos direccionais) ou de utilidade para a administración de medioambiente (indicadores de ozono, partículas,…) que utilizan con frecuencia a aproximación de datos funcionais en base a medicións continuas ó longo de cada día.

1. **Documento 1B-2: Estado da arte da investigación proposta**

Fai unhas décadas era impensable que o desenrolo económico puidera afectar tan negativamente á natureza como para chegar a representar un serio problema. É por isto que nos últimos anos os gobernos e as organizacións ecoloxistas en conxunción cos científicos buscan solucións para prever a degradación medioambiental. En particular as accións políticas para a protección do medioambiente obrigan ás empresas a desenrolar plans de control mediambiental de calidade de aire, auga e solo.

A Estatística xoga un papel moi importante á hora de predicir os niveis de polución que as industrias causan no seu entorno. Existen diversas técnicas, de tipo paramétrico e non paramétrico, para a predición de valores futuros. Dentro da clase dos modelos paramétricos para a predición da tendencia inclúense os clásicos ARMA, que engloban como casos particulares ós puramente autorregresivos (AR) e ós modelos de medias móbiles (MA) (Wei, 1990;Brockwell e Davis, 1987). Estes modelos foron xeneralizados ó caso multidimensional dando lugar ós modelos VARMA (Lütkelpohl, 2005) e, posteriormente, ós VECM (Engle e Granger, 1987; Lütkelpohl, 2005) que teñen en conta a posible relación de cointegración entre as series. Por outra banda, tamén se desenvolveron modelos non paramétricos para a estimación da tendencia como os modelos aditivos (Hastie e Tibshirani, 1990), modelos baseados en redes neuronais (Tsay, 2002) ou a estimación tipo núcleo (Hastie e Tibshirani, 1990).

Na predición de indicadores medioambientais, así como no análise de datos procedentes de diversas ciencias, como a Bioloxía, a Xeoloxía ou a Meteoroloxía, tamén pode ser interesante ter en conta datos de tipo direccional, como por exemplo a dirección do vento. A maioría dos avances teóricos na estatística de variables aleatorias direccionais é relativamente recente. A referencia chave na teoría desta área é o libro de Mardia (2000) así como o de Fisher (1993).

O progreso das ferramentas de computación, tanto en termos de memoria como de capacidade, permite traballar con grandes conxuntos de datos. En particular, é habitual atoparnos cunha variable aleatoria observada en continuo, polo que un pode chegar a considerar que o que observa é realmente a realización dunha variable funcional. Isto fixo que nos últimos anos os datos funcionais estean de plena actualidade, e revistas de máximo nivel como Statistica Sinica (Vol.14,Nº3,2004) ou Computational Statistics & Data Analysis (Vol.51,10,2007) publicaran números monográficos adicados a desenvolvementos metodolóxicos con datos funcionais e aplicacións. O libro que máis contribuíu a popularizar as técnicas estatísticas para datos funcionais é sen dúbida algunha o publicado por Ramsay e Silverman (2005) que considera o caso dos modelos de regresión con datos funcionais a través de procedementos próximos ó suavizado spline. O outro hito bibliográfico é o de Ferraty e Vieu (2006) que trata os modelos de regresión mediante a extensión da estimación tipo núcleo. Febrero (2010) trata as medidas de influencia do modelo funcional lineal con resposta escalar no que a previsión é funcional. Anteriormente xa se propuxeron varios enfoques para a estimación deste modelo, como por exemplo Cardot, Ferraty e Sarda (2003) ou Ramsay e Silverman (2005) que analizaron o uso de funcións de forma restrinxida e métodos de penalización. Recentemente deseñáronse modelos máis flexibles que permiten conxugar a estimación da compoñente funcional con outras covariables que non o son (Aneiros e Vieu, 2009; Baíllo e Grané, 2009). Paralelamente a esta evolución teórica neste campo tamén se logrou un gran avance na creación e optimización de novo software científico, sobre todo no que se refire o sotfware estatístico de libre distribución R. Indubidablemente a referencia básica para traballar en R é o paquete fda (Ramsay, 2010) e desde o punto de vista non paramétrico os autores do grupo STAPH manteñen activa a web <http://lsp.ups-tlse.fr/staph> co código R das súas funcións.

A lexislación de calidade de aire obriga ás industrias a controlar os niveis de contaminación que causan no seu entorno como se comentou ó inicio. En particular, na Central Térmica de As Pontes implantouse un sistema suplementario da contaminación atmosférica. Dito sistema sufriu diversas evolucións ó longo dos anos con diferentes modelos de predición de tipo semiparamétrico do SO2 dos que cabe destacar:

1. Un modelo de predición non paramétrico con erros ARI, García Jurado e outros (Technometrics, 1995)
2. Modelización de tipo semiparamétrico espacio-temporal, Angulo e outros (Environmental and Ecological Statistics, 1998)
3. Modelización con modelos parcialmente lineais, Prada Sánchez e outros (Environmetrics, 2000)
4. Modelización usando redes neuronais, Fernández de Castro e outros (Air & Waste Manage, 2003).
5. Predición con modelos de resposta binaria de tipo semiparamétrico, Roca Pardiñas e outros (Environmetrics, 2004).
6. Modelización con modelos funcionais, Fernández de Castro e outros (Technometrics, 2005)
7. Estimación e contrastes con modelos aditivos xeneralizados, Roca Pardiñas e outros (Statistics and Computing,2005)
8. Predición utilizando técnicas de aprendizaxe tipo boosting, Fernández de Castro e González Manteiga (Stoch. Environ. Res. Risk. Asses., 2007)

Na actualidade vívese un momento apaixonante dende o punto de vista científico. A Central Térmica sufriu un proceso de adaptación que permite o seu funcionamento con 100% carbón de importación, as emisións de SO2 reducíronse en torno a un 95% e empezou a funcionar unha nova central de ciclo combinado onde a serie de NOx cobra vital importancia. Estamos ante o desafío de predicir unha serie con resposta bidimensional onde os erros bidimensionais poden ter distinta estrutura paramétrica e teñamos por tanto que xogar con aspectos de cointegración , tratados en el contexto de la Economía pero escasamente en medioambiente. Na central xa está implantado un sistema que predí, con media hora de antelación, os niveis de SO2 e de NOx de forma independente. Estase traballando na modelización semiparamétrica con estrutura de correlación nos erros e os primeiros resultados xa foron presentados en congresos nacionais (SEIO 2009) e internacionais (ISI 2009). En esta última sesión por invitación en una sesión invitada dedicada a la predicción con indicadores de medioambiente liderada por…(darle mucha importancia a este párrafo). Ademais, deseñouse a primeira versión dunha ferramenta que permite decidir cal é a orixe dun episodio de alteración da calidade do aire: o ciclo combinado, a central térmica ou outros posibles focos de contaminación, na que xoga un papel moi importante unha covariable de tipo direccional: a dirección do vento. (PFM de Fran…)

O noso grupo de investigación tamén abordou outros problemas no eido medioambiental ademais do da central térmica. Por exemplo, Febrero e outros (2008) estudaron os niveis anómalos de NOx na cidade de Barcelona mediante medidas de profundidade funcional.

1. **Documento 1B-3: Obxectivos do proxecto**

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

1. Elaboración de metodoloxía dentro do campo dos modelos de series temporais
   1. Modelización semiparamétrica multidimensional con estrutura de correlación nos erros
      1. Modelización do erro con estrutura VAR.
      2. .Modelización do erro con estrutura VECM: relacións de cointegración.
   2. Estimación, contrastes e predición en ditos modelos.
   3. Intervalos de predición: técnicas clásicas e metodoloxía Bootstrap.
   4. Estudio de simulación e validación dos modelos anteriores.
   5. Aplicación dos modelos estudados nos apartados anteriores á predición de series medioambientais.
2. Elaboración de metodoloxía dentro do campo dos datos direccionais.
   1. Estudio de distribucións notables de datos direccionais.
   2. Estudio de modelos de predición con datos direccionais.
   3. Aplicación do estudado en 2.1 e 2.2 á predición de series medioambientais.
3. Elaboración de metodoloxía dentro do campo dos datos funcionais.
   1. Estudio do modelo de regresión con covariables funcionais e resposta real ou funcional.
      1. Modelo lineal xeneralizado (GLM)
      2. Modelo parcialmente lineal (LMP)
      3. Estudio aditivo xeneralizado (GAM)
      4. Clasificación
   2. Estudio de simulación e validación dos modelos anteriores.
   3. Aplicación dos modelos do primeiro apartado á predición de series temporais e espacio-temporais medioambientais con datos direccionais.
4. Elaboración de software
   1. Transferencia tecnolóxica da nova metodoloxía á Central Térmica de As Pontes.
   2. Creación e optimización do novo software científico creado a partir do estudio desenrolado nos apartados anteriores.

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.

(HABLAR MUCHO MAS DE LA TRANSFERENCIA)

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,do proxecto final do Máster en Técnicas Estatísticas que está a cursar unha das integrantes do equipo e das tese doutorais doutros membros do grupo.

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. HABLAR DEL CONGRESO METMA Y DEL IWFOS2011

Para a construción do software empregarase a linguaxe de programación R (software libre), ferramenta empregada para a creación dun paquete de datos funcionais que estará a disposición de futuros usuarios no repositorio de contribucións (http://cran.proyect.org). Ademais aproveitando que o departamento pertence ao programa CONSOLIDER, incluirase no repositorio de software que se ten previsto crear dentro dito programa. (REDACCION INCORRECTA) HABLAR DE LA RED:

Todos os resultados que se van obter permitirán o estudo de datos medioambientais ou de tipo ecolóxico, a predición da evolución de indicadores de polución,... 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 o principal 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. Outro objetivo es clasificar as fontes de inmisión dos niveles de SO2 y NOx para coñecer con maior precisión e rapidez si el orixe foi da Central, do ciclo combinado ou doutra fonte exóxeno a Central.

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 y con datos reales; 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. Aspecto que se viene repitiendo desde sus inicios.

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**

Tanto o IP Wenceslao González coma José M. Prada e Manuel Febrero, teñen ampla experiencia en diversas liñas de investigación: Inferencia en modelos espazo temporais (aplicación na predición medioambiental con mapas de SO2 e dirección de vento, predición xeolóxica, modelos ecolóxicos, etc.), Modelización econométrico-financeira, Técnicas de aprendizaxe (aplicación no control industrial, aplicacións medioambientais na predición de niveis de SO2, etc.), tests de bondade de axuste para modelos diversos, Inferencia con datos funcionais (aplicación á predición de series de tipo de interese, de curvas de ozono, etc.), … Ademais, tamén son integrantes de varias redes de investigación tanto a nivel nacional (Nodo CESGA, programa CONSOLIDER) como internacional (IAP Network), e do proxecto de investigación de convocatoria pública: “Inferencia Estadística funcional. Nuevas aplicaciones a las Finanzas, al Medioambiente, a la Industria y a la Economía” (2005-2008); IP: Wenceslao González Manteiga; subvencionado polo Ministerio de Educación y Ciencia. (FALTA PROYECTO ACTUAL Y LA RED) Por razóns de espazo omítense a participación en proxectos con empresas, capítulos de libros ou proceedings, e numerosas aportacións en distintos congresos de rango nacional e internacional. Outros membros do grupo, María Piñeiro, Adela Martínez e Manuel Oviedo, teñen experiencia na participación en outros proxectos de investigación a nivel autonómico e nacional como se pode consultar nos seus currículums. Ademais María Leyenda ten un contrato financiado por Endesa e realiza viaxes dúas veces a semana a central de As Pontes.

Só se citan algúns dos artigos máis importantes e recentes (para máis información consultar os currículums).

1. **Documento 1B-13: Instalacións e instrumentos**
2. **Documento 1B-14: outros medios**