**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 na 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. Arnold e SenGupta (2006) realizan unha revisión dos avances máis recentes na análise deste tipo de datos nas ciencias ecolóxicas e ambientais. Un exemplo aplicado é o tratado por Xu e Schoenberg (2007), onde se utiliza a regresión núcleo para investigar a relación entre unha variable explicativa de dirección e unha resposta de valor real variable. Aínda que a maioría dos avances teóricos na estatística de variables aleatorias direccionais é relativamente recente, a referencia chave é o libro de Mardia (2000) o igual que 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. A análise de compoñentes principais funcionais vense usando de forma habitual para estimar o parámetro do modelo lineal funcional, sen embargo no traballo de Martínez-Calvo (2008) proponse modificar esta metodoloxía mediante un presuavizado ben sobre a variable explicativa funcional, ben sobre a variable resposta escalar. 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). Na literatura abundan traballos que se adican á clasificación supervisada (James e Sugar, 2003; Yao e outros, 2005). Ferraty e Vieu (2006) resolven o problema de discriminación mediante técnicas non paramétricas (procedemento dos k veciños máis próximos ou estimadores tipo núcleo). Outros enfoques son os presentados por Cuevas e outros (2007) que se basean en conceptos de profundidade para clasificar o dato funcional e, Cuesta-Albertos e outros (2007) empregando proxeccións aleatorias. Un campo máis novidoso xurde da posibilidade de adaptar as técnicas funcionais ó contexto xeoestatístico. Giraldo e outros mostran como as técnicas de análise de datos funcionais poden estenderse ó campo espazo-temporal mediante técnicas kriging (Stein, 1999).

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 software 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 de control 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 espazo-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 , amplamente tratados no contexto da Economía pero con escasa aplicación ó 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). Neste último nunha sesión invitada adicada á predición con indicadores de polución e a súa repercusión na normativa de calidade de aire, liderada por Peter Guttorp, referencia imprescindible cando se fala de estatística medioambiental. Proba disto é que é membro do grupo intergubernamental sobre o cambio climático (Premio Novel da Paz do 2007) e codirector da revista Environmetrics. 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. Esta primeira versión foi o resultado dun proxecto de fin de máster dirixido por dous membros do noso equipo de investigación.

Toda a metodoloxía elaborada non só ten aplicación a este problema concreto, senón que poderá ser integrada nos sistemas de control da contaminación atmosférica en outras centrais térmicas que utilicen como combustible principal o carbón nacional.

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.

Por razóns de espazo omítense as referencias bibliográficas. Algunhas delas poden ser consultadas nos currículums dos membros do grupo.

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. Análise de distribucións notables de datos direccionais.
   2. Elaboración 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 (PLM)
      3. Modelo aditivo xeneralizado (GAM)
   2. Clasificación funcional supervisada:
      1. Métodos non paramétricos: estimador KNN  e estimador núcleo.
      2. Método Boosting.
   3. Estudio de simulación e validación dos modelos anteriores.
   4. Aplicación dos modelos do primeiro apartado á predición de series temporais e espazo-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**

Os datos direccionais encóntranse en tódalas áreas das ciencias aplicadas, aínda que máis frecuentemente en ecoloxía e ciencias do medio ambiente.Os datos direccionais son útiles á hora de desenrolar unha metodoloxía que nos permita realizar previsións precisas tanto de indicadores de polución coma predicións da enerxía eólica a longo prazo. A enerxía eólica forma parte das enerxías renovables e ten a vantaxe de que é inesgotable, limpa e segura. Sen embargo o gran inconveniente que presenta hoxe en día é o seu elevado erro de previsión para horizontes non moi lonxanos, debido á gran incertidumbre que supón a predición do vento.

En particular este proxecto, tratará de establecer distintos modelos de tipo paramétrico ou non paramétrico para modelos de regresión con covariable direccional e resposta continua ou binaria. Os modelos aplicaranse a datos de tipo medioambiental, onde a covariable fundamental é a dirección do vento e, a resposta son indicadores de niveis de dióxido de xofre ou óxidos de nitróxeno.

Un dos obxectivos finais deste proxecto é poder continuar ca 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 da que nace un sistema preciso de predición de indicadores de polución. Esta ferramenta denomínase Sistema de Predición Estatística de Inmisión (SIPEI) e sufriu múltiples adaptacións ó longo destes anos. A implantación e a súa mellora producíronse grazas ó compromiso de Endesa co medio ambiente e o seu interese por mellorar e adaptarse as distintas circunstancias. Este sistema de predición é unha importante ferramenta de información e axuda na toma de decisións na central.

O SIPEI emprega datos de inmisión minutais, dispoñibles a través do sistema MEDAS, para obter predicións dos valores de dióxido de xofre e óxidos de nitróxeno, con media hora de antelación, utilizando modelos aditivos. Este sistema de predición tamén proporciona unha predición probabilística de superación dos niveis de SO2 e NOx, tendo en conta os límites contemplados na lexislación vixente. Ademais, calcula a probabilidade estimada de que a central térmica ou o ciclo combinado sexan os causantes dun certo episodio de alteración da calidade do aire.

Nos últimos anos, o SIPEI 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 coa orixe, a estrutura e a maneira de obter os datos. Constitúe a unión entre o provedor dos datos e o propio sistema de predición. Neste módulo obtéñense os datos orixinais, formatéanse convenientemente e póñense a disposición do módulo de predición.
* A predición fai referencia ós aspectos máis metodolóxicos, matemáticos e estatísticos. Neste módulo encóntranse os procesos xurdidos da implementación dos modelos desenrolados. Ademais, lense os datos formateados, aplícanse os modelos, realízanse os cálculos, obtéñense os resultados e se poñen a disposición do módulo de visualización.
* A visualización comporta aspectos relacionados coa disposición e presentación da información. Neste módulo lense os resultados, xéranse os gráficos e os informes e ponse toda a información a disposición do usuario.

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

Os avances logrados durante a execución deste proxecto estarán ó 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 teses 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, Environmetric, Journal of Chemometrics, Computational Statistics, Data Analysis, Journal of Geophysical Research, Atmosphere, Wind energy, Journal of Multivariate analysis, Journal of Agricultural, Biological, and Environmental Statistics, Ecological Applications.…) ou de carácter máis computacional (Journal of Statistical Software). 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). O departamento tamén forma parte da 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. Ademais o grupo tamén pertence á rede matemática Consulting and Computing de Galicia, da que é coordinador o IP deste grupo. Este nodo xorde dentro do proxecto [Consolider Ingenio MATHEMATICA](http://www.i-math.org/) coa misión de tomar iniciativas encamiñadas á transferencia de coñecemento matemático ó sector produtivo, promovendo o uso de métodos e técnicas matemáticas na industria e na empresa en xeral, na xestión do medio ambiente, así como a investigación en temas de interese para o desenrolo tecnolóxico.

Recentemente o departamento organizou o V *International Workshop on Spatio-Temporal Modelling* (METMAV: <http://eio.usc.es/pub/metma/>) celebrado na facultade de Matemáticas de Santiago de Compostela desde o 30 de Xuño ó 2 de Xullo de 2010, promovendo así o desenrolo e a aplicación de métodos estatísticos espazo-temporais en diferentes campos relacionados coas Ciencias Medioambientais. Ademais, un número especial de [*Environmetrics*](http://www3.interscience.wiley.com/journal/6285/home) será editado cos traballos presentados e terá como editor xefe o IP da presente solicitude. Tamén se publicará outro número especial do [*Journal of Environmental Statistics*](http://www.jenvstat.org/).

Hai que sinalar ademais que a próxima edición do *International Workshop on Functional and Operatorial Statistics* (IWFOS) celebrarase en Santander (España) en 2011. O IP do noso grupo está a participar activamente na súa organización, posto que é coorganizador xunto con Federic Ferraty (Univ. de Toulouse) e Juan Cuesta (Univ. de Cantabria). Este evento é moi importante en canto a difusión de resultados xa que atraerá ós estatísticos interesados no desenrolo de métodos que afrontan o problema da elevada dimensionalidade.

Para a construción do software empregarase a linguaxe de programación R (software libre). Está ferramenta será empregada para a creación dun paquete de datos funcionais que estará a disposición de futuros usuarios no repositorio de contribucións (http://cran.proyect.org).

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 existe un amplo abanico de liñas de investigación que aínda se atopa sen explorar.

Por iso un dos principais obxectivos deste proxecto será estender os modelos de predición xa construídos no ámbito unidimensional e multidimensional ó terreo funcional, o que é de grande interese desde o punto de vista teórico. Outro obxectivo será clasificar as fontes de inmisión dos niveles de SO2 e NOx para coñecer con maior precisión e rapidez se a orixe dun episodio de alteración de calidade de aire é a central térmica, o ciclo combinado ou outra fonte esóxena.

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 ó campo das series de tempo, o que é de gran interese tanto desde un punto de vista teórico como práctico. Ademais, isto será beneficioso á hora de construír modelos de predición xa que se poderán incluír variables meteorolóxicas (por exemplo, dirección e velocidade do vento) ó longo do tempo e perfeccionar así os sistemas de predición ambientais.

Cabe notar que estes obxectivos están orientados nun principio a ser aplicados na U.P.T As Pontes, aínda que tamén estes resultados serán aplicables noutras empresas, como por exemplo en centrais térmicas que empreguen como combustible principal carbón nacional.

A xeneralidade e validez dos modelos de predición implementados empregando datos funcionais e direccionais, permitirán estendelos tanto a outros sectores enerxéticos (xeración de enerxía eólica) como sectores medioambientais (control de incendios, indicadores de polución,…).

1. **Documento 1B-7: Metodoloxía**

No que ó desenvolvemento matemático se refire, a metodoloxía a seguir será a habitual nestes casos empregando tanto ferramentas estatísticas como probabilísticas. Primeiramente, se desenrolará nova metodoloxía nos tres grandes campos de estudo: deseño dun modelo semiparámetrico multidimensional no eido das series temporais, elaboración dun modelo de predición con datos direccionais e construción dun novo estimador para a regresión funcional. Seguidamente, se levará a cabo a validación de tódolos modelos a través de estudos de simulación e aplicación os datos reais. Ademais se fará unha extensión das técnicas de aprendizaxe ós datos funcionais.

En canto ó aspecto computacional, o software que se constrúa estará escrito na linguaxe de programación R. O entorno de programación estatística R consolidouse como unha plataforma de desenrolo de algoritmos estatísticos avanzados. É de libre distribución e gratuíto, o que fai accesible a toda a comunidade científica, académica e profesional os resultados das aplicacións e métodos implementados. As técnicas e tratamentos básicos xa están implementados no repositorio de contribucións (<http://cran.proyect.org>). Ademais, ó 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 a integrante do equipo María Leyenda realiza dúas veces á semana a Central Térmica de As Pontes, permitirán a transferencia tecnolóxica dos traballos realizados á área medioambiental. Cómpre salientar a relación entre o grupo e a Sección de Medio Ambiente da Central Térmica de As Pontes desde fai 22 anos (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 ó 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, e agora de NOx, como se amosa no diagrama que se anexa.

1. **Documento 1B-8: Plan de traballo**

O traballo do proxecto está planificado para unha duración de tres anos. A continuación detállanse as actividades a realizar en cada un deles.

Primeiro ano.

Levarase a cabo unha análise exploratoria e unha profunda revisión das metodoloxías relativas ás series temporais (con especial atención as multidimensionais e a cointegración), datos funcionais e datos direccionais. A partir de dito traballo se desenrolaran os modelos propostos nos obxectivos 1.1, 2.1 e 3.1, seguido da estimación e contrastes dos correspondentes modelos.

Segundo ano.

Traballarase con técnicas de remostraxe para a construción dos intervalos de predición dos modelos 1.1 mediante técnicas Bootstrap e realizaranse simulacións para a súa validación. Ademais, construiranse diversos modelos de clasificación supervisada empregando técnicas funcionais. Isto corresponde ó cumprimento dos obxectivos 1.3 a 1.4 e 2.2. Tamén se desenvolverán modelos de predición con datos direccionais (obxectivo 3.2).

Se realizarán estudos de simulación e validación dos algoritmos desenrolados, que estarán a disposición da comunidade científica a través da web do proxecto. Se desenrolará un modulo de análise exploratorio e outro de modelos de regresión para datos funcionais que formarán parte dunha librería de R. Dita librería se colgará no repositorio de contribucións (<http://cran.proyect.org>). Isto se corresponde co punto 4.2.

Terceiro ano.

Se realizaran estudos de simulación e validación dos algoritmos desenrolados (obxectivo 1.4 e 3.3) e se colgarán na web do proxecto. Aplicaranse os resultados obtidos durante os dous anos anteriores a datos medioambientais (obxectivos 1.5, 2.3 e 3.4). O software construído para o ámbito funcional engadirase no repositorio de contribucións (<http://cran.proyect.org>), dito software favorecerá a transferencia tecnolóxica a empresas como ENDESA. Isto corresponde ó cumprimento do obxectivo 4.1.

Tódolos membros do equipo participarán activamente nas tarefas anteriormente mencionadas. Por outra banda, ó longo de todo o proxecto haberá reunións periódicas con representantes de ENDESA, o que axudará a mellorar a aplicabilidade dos traballos realizados.

Anéxase un cronograma coas tarefas previstas para os tres anos de duración do proxecto.

1. **Documento 1B-11: Explicación e xustificación dos gastos**
2. **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, …), Modelización econométrico-financeira, Técnicas de aprendizaxe (aplicación no control industrial, aplicacións medioambientais na predición de niveis de SO2, …), 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 (proxecto [Consolider I-MATH](http://www.i-math.org/) dentro da rede matemática Consulting; nodo Cesga) como internacional (IAP Network), e dos proxectos de investigación de convocatoria pública: “Metodología y aplicaciones en Estad´sitica Semiparamétrica” (2005-2008) e “[Metodologías y aplicaciones en estadística semiparamétrica, funcional y espazo-temporal”](http://eio.usc.es/files/PROYECTO_FICHA.php?lan=es&Id_proyecto=124) (2009-2013); con IP: Wenceslao González-Manteiga; subvencionados polo Ministerio de Educación y Ciencia, do que tamén forman parte Adela Martínez e María Piñeiro. 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.

Por razón de espazo só se citan as 18 máis importantes e recentes das 78 publicacións dende 2005 e relacionadas co proxecto. Para máis información consultar os currículums.

1) Cuesta-Albertos, JA., Febrero-Bande, M. (To appear). "A simple multiway ANOVA for functional data". TEST. Springer

2) Martínez Calvo, A., González-Manteiga, W. (To appear). "Bootstrap in functional linear regression". Journal of statistical planning and inference. Elsevier Science

3) Reboredo, J.C., González-Manteiga, W., Febrero-Bande, M., Matías, J.M. (2010). "Boosting GARCH and neural networks for the prediction of heteroskedastic time series". Mathematical and Computer Modelling. 51. 256-271. Elsevier

4) Crujeiras, R., Fernández-Casal, R., González-Manteiga, W. (2010). "Goodness-of-fit tests for the spatial spectral density". Stochastic Environmental Research and Risk Assesment. 24. pp. 67-79. Springer

5) Febrero-Bande, M., Galeano, P., González-Manteiga, W. (2010). "Measures of influence for the functional linear model with scalar response". Journal of Multivariate Analysis. 101. 327-339. Elsevier

6) González-Manteiga, W., Ordoñez, C., Taboada, J., Matías, J.M. (2010). "Partially linear support vector machines applied to the prediction of mine slope movements". Mathematical and Computer Modelling. 51. pp. 206-215. Elsevier

7) Castellano, M., Franco, A., Cartelle, D., Febrero-Bande, M., Roca, E, (2009). "Identification of NOx and Ozone Episodes and Estimation of Ozone by Statistical Analysis". Water, Air and Soil Pollution. 198. 95–110

8) Crujeiras, R., Fernández-Casal, R., González-Manteiga, W. (2009). "Nonparametric test for separability of spatio-temporal processes". Environmetrics. DOI\_10.1002/env.1006. Wiley

9) Menezes, R., García-Soidán, P.H., Febrero-Bande, M. (2008). "A kernel variogram estimator for clustered data". Scandinavian Journal of Statistics. 35, 1. 18-37. Blackwell Synergy10) Febrero-Bande, M., Galeano, P., González-Manteiga, W. (2008). "Outlier detection in functional data by depth measures with application to identify abnormal NOx levels". Environmetrics. 19, 4. 331-345

11) Fernández de Castro, B.M., González-Manteiga, W. (2008). "Boosting for real and functional samples. An application to an environmental problem". Stochastic Environmental Research adn Risk Assesment. 22, 1. 27-37

12) Martinez, M.D., Raya-Miranda, R., González-Manteiga, W., González-Carmona, A. (2008). "A bootstrap local bandwidth selector for additive models". Journal of Computational and Graphical Statistics. 17, 1, 38-55

13) Raya-Miranda, R., Martinez, M.D., González-Manteiga, W. (2008). "Sizer map for inference with additive models". Statistics and Computing. 18. pp. 297-312. Springer

14) Crujeiras, R., Fernández-Casal, R., González-Manteiga, W. (2007). "Comparing spatial dependence structures using spectral density estimators". Environmetrics. 18. pp. 793-808. Wiley InterScience

15) Febrero-Bande, M., Cuevas, A., Fraiman, R. (2006). "On the use of the bootstrap for estimating functions with functional data". Computational Statistics & Data Analysis. 51, nº 2. 1063-1074.16) Fernández de Castro, B.M., Guillas, S., González-Manteiga, W. (2005). "Functional Samples and Bootstrap for Predicting SO2 Levels". Technometrics. 47, n. 2. 212-222

17) Roca, J., Cadarso, C.M., González-Manteiga, W. (2005). "Testing for interactions in Generalized Additive Models:application to SO2 pollution". Statistics and Computing, 15, 289-299.

|  |
| --- |
| 18) Aira, M.J., Castellano, M., Iglesias, I., Jato, V., González-Manteiga, W. (2005). "Artificial neural networks as a useful tool to predict the risk level of Betula pollen in the air". International journal of Biometeorology. 49. 310-316. Springer |

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