



UNIVERSIDAD DE CÓRDOBA



**ESTIMACIÓN DE LOS FLUJOS DE REMESAS ASOCIADOS A
CORREDORES MIGRATORIOS EN CONDICIONES DE
INCERTIDUMBRE: EL CASO DE ECUADOR**

TESIS QUE PRESENTA

MARÍA DEL PILAR CAMPOY MUÑOZ

PARA OPTAR AL GRADO DE DOCTOR POR LA UNIVERSIDAD DE CÓRDOBA

CÓRDOBA, JUNIO DE 2013

DIRECTORES

CARLOS R. GARCÍA ALONSO

MELANIA SALAZAR ORDÓÑEZ

PEDRO ANTONIO GUTIÉRREZ PEÑA

CENTRO CIENCIAS ECONÓMICAS Y EMPRESARIALES ETEA

(CENTRO ADSCRITO A LA UNIVERSIDAD DE CÓRDOBA)

TITULO: *ESTIMACIÓN DE LOS FLUJOS DE REMESAS ASOCIADOS A
CORREDORES MIGRATORIOS EN CONDICIONES DE
INCERTIDUMBRE: EL CASO DE ECUADOR*

AUTOR: *MARÍA DEL PILAR CAMPOY MUÑOZ*

© Edita: Servicio de Publicaciones de la Universidad de Córdoba.
Campus de Rabanales
Ctra. Nacional IV, Km. 396 A
14071 Córdoba

www.uco.es/publicaciones
publicaciones@uco.es

La tesis titulada “Estimación de los flujos de remesas asociados a corredores migratorios en condiciones de incertidumbre: el caso de Ecuador”, que presenta D^a María del Pilar Campoy Muñoz para optar al grado de Doctor, ha sido realizada dentro del programa de Doctorado “Ciencias Jurídicas y Sociales” del Departamento de Gestión Empresarial y Métodos Cuantitativos de la Universidad de Córdoba (ETEA) bajo la dirección del Doctor D. Carlos R. García Alonso, del Doctor D. Melania Salazar Ordoñez y del Doctor D. Pedro Antonio Gutiérrez Peña.

En Córdoba, junio de 2013

El Doctorando

El Director



Fdo: María Pilar Campoy Muñoz

Fdo: Carlos R. García Alonso

El co-director

El co-director

Fdo: Melania Salazar Ordóñez

Fdo: Pedro Antonio Gutiérrez Peña

Esta tesis doctoral ha sido financiada en parte con cargo a los Proyectos **TIN2011-22794** de la Comisión Interministerial del Ministerio de Ciencia y Tecnología (MICYT), al Proyecto de Excelencia **P2011-TIC-7508** de la Junta de Andalucía y con fondos FEDER de la Unión Europea.



TÍTULO DE LA TESIS:

ESTIMACIÓN DE LOS FLUJOS DE REMESAS ASOCIADOS A CORREDORES MIGRATORIOS EN CONDICIONES DE INCERTIDUMBRE: EL CASO DE ECUADOR

DOCTORANDO/A:

MARIA DEL PILAR CAMPOY MUÑOZ

INFORME RAZONADO DEL/DE LOS DIRECTOR/ES DE LA TESIS

(se hará mención a la evolución y desarrollo de la tesis, así como a trabajos y publicaciones derivados de la misma).

La Tesis Doctoral titulada:

ESTIMACIÓN DE LOS FLUJOS DE REMESAS ASOCIADOS A CORREDORES MIGRATORIOS EN CONDICIONES DE INCERTIDUMBRE: EL CASO DE ECUADOR

es la culminación de un proceso intensivo de investigación científica sobre el desarrollo de metodologías aplicables a la modelización y realización de pronósticos relativos a fenómenos socio-económicos complejos en situaciones de incertidumbre, como ocurre en el caso de los flujos de remesas. Durante el desarrollo de esta investigación, la doctoranda ha asistido a eventos científicos de relevancia nacional e internacional y publicado artículos en revistas indexadas. El proceso de investigación ha incluido: i) la revisión sistemática e intensiva de la bibliografía científica relacionada con la materia objeto del estudio, ii) el diseño y desarrollo de una metodología híbrida, empleando para ello técnicas estadísticas y de inteligencia artificial, iii) el diseño y desarrollo de herramientas informáticas para la resolución de los distintos problemas matemático-estadísticos planteados y, finalmente iv) la validación de las distintas metodologías utilizando tanto el conocimiento experto como información secundaria disponible sobre el caso de estudio elegido.

La presente Tesis Doctoral es resultado de un proceso de colaboración entre diferentes grupos de investigación, siendo en nuestra opinión, un ejemplo de investigación científica que trasciende el mero conocimiento académico y se proyecta sobre la planificación y gestión de situaciones reales. Teniendo en cuenta esta característica, las principales dificultades que se afrontaron fueron: i) la adaptación de las metodologías de análisis matemático y estadístico para la correcta manipulación de los distintos tipos de incertidumbre existentes en fenómenos complejos; ii) la dificultad teórica que existe en este ámbito científico; y, finalmente iii) la existencia de un conocimiento muy fragmentado sobre el fenómeno objeto de estudio que hizo especialmente ardua la tarea del diseño de la base de conocimiento sobre la que opera la metodología diseñada. Todas estas dificultades se han abordado con rigurosidad científica y validado mediante la estimación de los flujos de remesas, en el corredor España-Ecuador durante el periodo 2000-2010.

Las metodologías utilizadas en este trabajo de investigación pueden ser de especial utilidad en el ámbito del estudio y la planificación de políticas públicas relativas al desarrollo económico de los países receptores de flujos de remesas. Teniendo en cuenta la evidencia de los resultados obtenidos, los entes decisores tienen a su disposición un conjunto flexible de herramientas para apoyar sus decisiones en este marco de referencia.

En la actualidad, el equipo de investigación del cual forma parte esencial la doctoranda continúa trabajando en estos temas, existiendo la posibilidad cierta de extender la investigación utilizando nuevas metodologías de análisis y ampliando el ámbito de aplicación de la herramienta diseñada a otros problemas socio-económicos.

Por todo ello consideramos que la Tesis Doctoral presentada tiene todos los avales necesarios para su presentación y defensa.

Congresos:

Campoy-Muñoz, P.; García-Alonso, C.; Salazar-Ordoñez, M. (Junio 2013): Dealing with migratory flows under uncertainty: a multi-objective optimization approach. 22nd International Conference on Multiple Criteria Decision Making (Málaga, España).

Campoy-Muñoz, P.; Gutiérrez-Peña, P.A.; Hervás-Martínez, C. (Junio 2013): Addressing remitting behavior using an ordinal classification approach. 5th. International Work-Conference on the interplay between natural and artificial computation (Mallorca, España).

Campoy-Muñoz, P.; Salazar-Ordoñez, M.; García-Alonso (Mayo 2012): Remittance flows estimation hybridizing a multilevel fuzzy system and a simulation model. 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD'12) (Chongqing, China).

Campoy-Muñoz, P.; Salazar-Ordoñez, M.; García-Alonso (Octubre de 2011) Comunicación: A bayesian model for remittances and gender: the Spain-Ecuador corridor. International Conference "New debates on gender, migration and development: linking production and reproduction chains in contemporary mobility" (Menorca, España).

Campoy-Muñoz, P.; Salazar-Ordoñez, M.; García-Alonso (Junio de 2011): Aproximación a un modelo causal para la estimación de remesas: factores determinantes y sus relaciones. III International Conference about Migrations. (Oporto, Portugal).

Artículos

García-Alonso, C.; Campoy-Muñoz, P.; Salazar-Ordoñez, M. A Multi-Objective Evolutionary Algorithm for enhancing Bayesian Networks hybrid-based modeling. Computers and Mathematics with Applications.

DOI:<http://dx.doi.org/10.1016/j.camwa.2013.01.029>. Índice de Impacto: 1,747

Capítulos de Libros

Campoy-Muñoz, P.; Salazar-Ordoñez, M.; García-Alonso, C. A bayesian network for remittances and gender in the Spain-Ecuador corridor. The International Handbook on

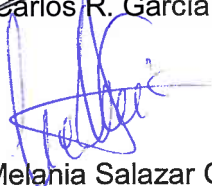
Gender, Migration and Transnationalism: Global and Development Perspectives.
Edwar Elgar Publisher (En imprenta).

Por todo ello, se autoriza la presentación de la tesis doctoral.
Córdoba, 3 de Junio de 2013

Firma del/ de los directores



Fdo.: Carlos R. García Alonso



Fdo.: Melania Salazar Ordoñez



Fdo.: Pedro Antonio Gutiérrez Peña

A mi padre.

A Juan.

AGRADECIMIENTOS

Afortunadamente, son muchos los que, directa o indirectamente, han contribuido a la realización de esta investigación y a los que deseo agradecer su ayuda.

En primer lugar a mis Directores, al Dr. Carlos García por su apoyo, dedicación y generosidad durante la realización de esta investigación, a la Dra. Melania Salazar por su sentido crítico y perfeccionista, de gran ayuda a lo largo de esta etapa de mi formación, y al Dr. Pedro Antonio Gutiérrez por su disponibilidad y amabilidad, atendiéndome siempre que lo he necesitado.

Agradezco también la ayuda y colaboración prestada por los profesores de la Universidad Loyola Andalucía, en especial, a César Hervás quien ha tenido la enorme generosidad de transmitirme su saber hacer y su pasión por la investigación.

Debo extender mi gratitud a todos mis compañeros de investigación en Córdoba y Sevilla, en especial, a Mónica de la Paz, ejemplo de superación y buen humor, a M^a Luisa Rodero, por su apoyo en todo momento y sus consejos siempre bien acertados, y a los miembros del grupo de investigación AYRNA de la Universidad de Córdoba.

Finalmente, quisiera agradecer a toda mi familia y amigos su constante aliento, cariño y comprensión. A mis padres, por su ejemplo, amor y apoyo incondicional, y especialmente a Juan, por ser como es, por lo que compartimos y lo que significa para mí.

ÍNDICE GENERAL

ÍNDICE DE TABLAS.....	19
ÍNDICE DE FIGURAS.....	21
I. INTRODUCCIÓN.....	23
1.1. Introducción a la propuesta metodológica	25
1.2. Introducción al estudio de los flujos de remesas.....	30
1.3. Estructura de la investigación	33
II. HIPÓTESIS Y OBJETIVOS	35
2.1. Hipótesis.....	37
2.2. Objetivos generales e instrumentales.....	37
III. METODOLOGÍA	41
3.1. La Teoría de la Información Generalizada.....	43
3.2. Axiomas para la formalización de la incertidumbre en la teoría T	45
3.2.1. Determinación de la estructura de las variables aleatorias de la red bayesiana	46
3.2.2. Definición de las relaciones causales de la red bayesiana.....	47
3.2.2.1. Modelado difuso de relaciones causales en redes bayesianas	48
3.2.3. Identificación de las variables y sus relaciones causales en el caso de los flujos de remesas	53

3.3. Las redes bayesianas como metodología para la representación de la incertidumbre en la teoría T	60
3.3.1. Elementos de una red bayesiana	60
3.3.2. Redes bayesianas y propiedades de Markov	62
3.3.3. Métodos para resolver probabilísticamente una red bayesiana	63
3.3.4. Construcción de redes bayesianas causales y evaluación de intervenciones.....	64
3.4. La propuesta metodológica como procedimiento de cálculo de la teoría T ...	66
3.4.1. El motor de simulación Monte Carlo para la evaluación de relaciones causales	66
3.4.2. El motor de inferencia difusa para la evaluación de relaciones causales difusas	70
3.4.2.1. Definición y funcionamiento de un motor de inferencia difusa convencional.....	70
3.4.2.2. Limitaciones en el diseño de un MID	73
3.4.2.3. Diseño de un motor de inferencia difusa para la evaluación de relaciones causales en la red bayesiana (MID-RB).....	74
3.4.2.3.1. Automatización del proceso de creación y evaluación de reglas: Parámetros del MID-RB	74
3.4.2.3.2. Funcionamiento del MID-RB	76
3.4.3. Validación de los resultados.....	79
3.4.4. El Algoritmo Evolutivo Multi-objetivo para la determinación de los pesos de las variables input en las relaciones difusas	79
3.4.4.1. Elementos de un Algoritmo Evolutivo Multi-objetivo	79

3.4.4.2. Funcionamiento de un AEMO	81
3.4.4.3. Diseño de un AEMO para la estimación de pesos en las relaciones causales difusas de la red bayesiana (AEMO-RB).....	85
3.4.4.3.1.Planteamiento del problema multi-objetivo en las relaciones causales difusas	85
3.4.4.4. Diseño del AEMO-RB.....	87
3.5. La entropía de Shannon como medida de la incertidumbre en la teoría T	90
3.6. La teoría T sobre la incertidumbre: una visión de conjunto	91
IV. RESULTADOS.....	95
4.1. Estudio 1. Aplicación del procedimiento híbrido simplificado a la aproximación de los flujos de remesas en el corredor España–Ecuador	97
4.1.1. La red bayesiana causal para la estimación de los flujos de remesas	97
4.1.2. Diseño experimental	99
4.1.2.1. Caso de estudio	99
4.1.2.2. Parámetros del procedimiento de cálculo	102
4.1.2.2.1. Estructura de las variables aleatorias	102
4.1.2.2.2. Parámetros para la evaluación de relaciones difusas	108
4.1.3. Caracterización del proceso de simulación.....	112
4.1.4. Estimaciones de flujos de remesas y validación con datos reales.....	114
4.1.5. Distribución de probabilidad conjunta y Entropía de Shannon.....	121
4.2. Estudio 2. Aplicación del procedimiento híbrido mejorado a la aproximación del tamaño de la población inmigrante ecuatoriana en España	123
4.2.1. La red bayesiana causal para la estimación del tamaño de la población	

inmigrante en el país de acogida	123
4.2.2. Diseño experimental	124
4.2.2.1. Caso de estudio	124
4.2.2.2. Parámetros del MSMC	124
4.2.2.3. Parámetros del AEMO-RB	126
4.2.3. Estimación de pesos en las relaciones causales difusas mediante el AEMO-RB	128
V. DISCUSIÓN	133
5.1. El procedimiento de cálculo de la teoría T sobre la incertidumbre	135
5.2. La operatividad de la teoría T : los estudios sobre flujos de remesas y población inmigrante en el corredor España-Ecuador	138
5.3. Difusión de la propuesta metodológica	143
VI. CONCLUSIONES Y LÍNEAS FUTURAS DE INVESTIGACION	145
6.1. Conclusiones	147
6.2. Líneas futuras de desarrollo de la investigación	150
REFERENCIAS BIBLIOGRÁFICAS	151
ANEXO I. PUBLICACIONES	167

ÍNDICE DE TABLAS

Tabla 1. Conjunto de reglas que definen una relación causal difusa	52
Tabla 2. Relaciones causales de la RB.....	99
Tabla 3. Características de los inputs de la RB	103
Tabla 4. Variación de la pendiente de la curva para las variables del modelo con comportamiento no lineal durante el periodo de estudio.....	107
Tabla 5. Relaciones de dependencia difusa: características de las variables inputs....	109
Tabla 6. Relaciones de dependencia difusas: características de las variables outputs	111
Tabla 7. Descriptivos sobre la medida de variabilidad del MSMC	113
Tabla 8. Descriptivos para los flujos de remesas estimados en el corredor España-Ecuador (Millones de dólares).....	115
Tabla 9. Entropía de Shannon.....	121
Tabla 10. Probabilidad de los flujos de remesas estimados para el periodo de estudio	122
Tabla 11. Relaciones causales de la sección <i>d-separada</i> de la red bayesiana	124
Tabla 12. Características de las variables raíz de la sección <i>d-separada</i>	125
Tabla 13. Relaciones de dependencia difusa: características de las variables inputs .	125
Tabla 14. Relaciones de dependencia difusas: características de las variables outputs	126
Tabla 15. Experimentos con el AEMO-RB: pesos y número de generaciones	128
Tabla 16. Descriptivos para los pesos (w_i) de las variables input (DW_t , SB_t y MC_t) en la relación difusa obtenidos por AEMO-RB.....	129

Tabla 17. Descriptivos para Nm_t (año 2000) 129

Tabla 18. Descriptivos para las funciones estandarizadas [0.1, 0.9] de las soluciones eficientes 131

ÍNDICE DE FIGURAS

Figura 1. Esquema metodológico de la investigación	39
Figura 2. Conjuntos difusos definidos sobre la variable aleatoria “coste de envío”	50
Figura 3. Estructura de un MID tipo Mamdani.....	71
Figura 4. Sistema de Inferencia Difusa de Mamdani.....	73
Figura 5. Funcionamiento de un AEMO	84
Figura 6. Procedimiento del AEMO-RB.....	88
Figura 7. Esquema resumen de la propuesta metodológica.....	92
Figura 8. Red bayesiana causal para la estimación de los flujos de remesas	98
Figura 9. Evolución de las remesas recibidas en Ecuador por país de origen.....	101
Figura 10. Flujo de remesas en el corredor España-Ecuador durante el periodo 2000-2010 (millones de dólares): valores registrados (línea discontinua) <i>versus</i> valores estimados (línea continua)	114
Figura 11. Valores estimados para las variables de la red relativas a la propensión media a remitir de los inmigrantes ecuatorianos en España	117
Figura 12. Valores estimados para las variables de la red relativas al perfil socio-demográfico y laboral del inmigrante ecuatoriano en España	118
Figura 13. Valores estimados para las variables de la red relativas al salario medio del inmigrante ecuatoriano en España	119
Figura 14. Valores estimados para las variables de la red relativas al tamaño medio del stock de población inmigrante ecuatoriana en España	120
Figura 15. Sección <i>d-separada</i> de la red bayesiana causal para la estimación del tamaño de la población inmigrante en el país de acogida.....	123

Figura 16. Stock de población migrante ecuatoriana en España (Año 2000): valores estimados y valores reales	130
Figura 17. Evolución de la función de aptitud FF para los distintos experimentos	131

I. INTRODUCCIÓN

1.1. Introducción a la propuesta metodológica

La Teoría de la Información Generalizada (Klir, 1991) se fundamenta en la reducción de la incertidumbre presente en situaciones decisionales mediante el uso del conocimiento. Este paradigma promueve el desarrollo de teorías que expresen formalmente la incertidumbre y permitan cuantificar la reducción de la misma, ocasionada por la generación de información útil para la toma de decisiones.

Según Klir (2006), el desarrollo completo de una teoría T sobre la incertidumbre debe contemplar cuatro condiciones fundamentales como son: i) la caracterización, mediante los axiomas adecuados, de los distintos tipos de incertidumbre existentes en la situación decisional; ii) la utilización de una metodología que permita modelar dicha situación; iii) el desarrollo de un procedimiento de cálculo que posibilite obtener información sobre la situación representada por el modelo; y iv) la utilización de una medida funcional de la incertidumbre.

En este marco teórico, la presente investigación pretende desarrollar una teoría T sobre la incertidumbre, empleando redes bayesianas construidas en base al conocimiento como metodología de modelado de situaciones en condiciones de incertidumbre. Las redes bayesianas (Pearl, 1985) son modelos gráficos probabilísticos basados en la idea de causalidad, ampliamente utilizados para la modelización de fenómenos complejos donde existe un sólido corpus teórico, con independencia de la ausencia o escasez de datos sobre dichos fenómenos. Así pues, las redes bayesianas han sido utilizadas en la modelización de enfermedades (Park y Cho, 2012), del comportamiento del consumidor (Kisioglu y Topcu, 2011), de perfiles criminales (Baumgartner *et al.*, 2008) o en la predicción del precio de activos financieros (Zuo y Kita, 2012).

En los supuestos de falta de datos, las redes bayesianas son construidas y evaluadas a partir del conocimiento elicitado de los expertos o recogido en fuentes de

información secundaria (Druzdzel y van der Gaag, 2000). Esto permite diseñar marcos lógicos y holísticos que facilitan el razonamiento sobre el fenómeno objeto de estudio, al tiempo que codifican este conocimiento de forma que resulta útil para la toma de decisiones en distintos ámbitos de la realidad económica y social tales como el diagnóstico clínico (Visscher *et al.*, 2009), la gestión económica (Hunter *et al.*, 2009) o la gestión de recursos naturales (Bromley *et al.*, 2005).

En la teoría T propuesta en esta investigación, el procedimiento de cálculo se ha diseñado a partir de la hibridación de técnicas estadísticas y de inteligencia artificial, concretamente: Simulación Monte Carlo, Lógica Difusa y Algoritmos Evolutivos Multi-objetivo, ofreciendo una aproximación novedosa en la evaluación de redes bayesianas construidas en base a conocimiento.

La Simulación Monte Carlo (Metropolis y Ulam, 1949) es una técnica que permite la obtención de muestras artificiales de datos a partir de un generador de números pseudo-aleatorios y de la correspondiente función de probabilidad, mediante la cual se caracterizan las variables aleatorias existentes en un modelo (Raychaudhuri, 2008). Este tipo de simulación es la base de distintos algoritmos (Cano *et al.*, 1996; Fung y Chang, 1990) empleados para la propagación de información en redes bayesianas, en las que el tamaño de las mismas convierte esta tarea en un problema NP-complejo (Cooper, 1990).

La Lógica Difusa (Zadeh, 1975) es un sistema de conceptos, principios y métodos que permiten formalizar el razonamiento humano expresado mediante el lenguaje natural, siendo una aplicación de la Teoría de los Conjuntos Difusos (Celikyilmaz y Türksen, 2009). El modelado difuso ha sido empleado en la definición de las probabilidades de eventos en redes bayesianas (Liu *et al.*, 2009; Penz *et al.*, 2012). Además de ser una técnica ampliamente utilizada para el modelado en condiciones de incertidumbre (Bosma *et al.*, 2010; Rizzi *et al.*, 2003; Shepherd y Shi, 2006), ya que permite trabajar con funciones no lineales y evita tener que realizar presunciones

estadísticas sobre las características de los datos y sus distribuciones de probabilidad (Pedrycz y Reformat, 1997).

Los Algoritmos Evolutivos Multi-objetivo (Schaffer, 1985) son herramientas destinadas a la resolución de problemas de optimización complejos que permiten la aproximación del conjunto de soluciones Pareto eficientes (Coello *et al.*, 2007). En el ámbito de la modelización de redes bayesianas, los Algoritmos Evolutivos han sido aplicados con éxito en entornos como el aprendizaje de redes bayesianas y en inferencia (Larrañaga *et al.*, 2013; Wong y Guo, 2008; Zhu *et al.*, 2011).

La hibridación de las técnicas anteriores en el procedimiento de cálculo de la teoría T permite salvar dos de las limitaciones que afectan al modelado y resolución de redes bayesianas construidas a partir del conocimiento, tales como la dificultad para la asignación de probabilidades a las variables aleatorias de la red (Renooij y Witteman, 1999) y el tratamiento de aquellas de naturaleza continua (Cobb *et al.*, 2005). Además, la hibridación de técnicas hace posible la formalización y evaluación de distintos tipos de incertidumbre que pueden existir de forma simultánea en problemas complejos. Mientras que la Simulación Monte Carlo permite formalizar la incertidumbre ocasionada por falta o escasez de datos, la Lógica Difusa posibilita el tratamiento de la incertidumbre motivada por la imprecisión del lenguaje natural con el que se expresa el conocimiento sobre el fenómeno de estudio. Por su parte, los Algoritmos Evolutivos Multi-Objetivo reducen la subjetividad del proceso de calibración del modelo.

Así pues, en la teoría T propuesta, una vez diseñado el modelo gráfico que representa el comportamiento del fenómeno objeto de estudio, las variables aleatorias, a menudo continuas, son caracterizadas mediante las correspondientes distribuciones estadísticas, completando así el modelo causal o red bayesiana. Las relaciones causales entre las variables son modeladas bien mediante la asignación de una forma funcional o algebraica identificada en la literatura, bien mediante la definición de reglas o proposiciones condicionadas difusas, en aquellos casos en los

que la relación causal existe pero sólo ha podido ser descrita en términos semánticos. De esta forma, en la red bayesiana diseñada, las relaciones casuales puramente probabilísticas pueden ser parcial o totalmente sustituidas por relaciones causales de carácter algebraico y difuso, respectivamente. Finalmente, la distribución de probabilidad conjunta P de la red bayesiana es estimada mediante la hibridación de técnicas estadísticas como la Simulación Monte Carlo y la Inferencia Difusa. Esta metodología permite explorar todo el espacio de valores de las variables aleatorias discretas o continuas del modelo, así como evaluar las relaciones casuales definidas de carácter probabilístico, algebraico o difuso.

Al igual que cualquier metodología basada en el uso de conocimiento experto, la teoría T propuesta no está exenta de un cierto grado de subjetividad, ya que dicho conocimiento se emplea para definir la estructura del modelo causal, las distribuciones estadísticas asociadas a las variables aleatorias del modelo y el conjunto de parámetros necesarios para la construcción y evaluación automática de las reglas difusas. Estos últimos son de dos tipos, aquellos relativos a la estructura de la regla difusa, y aquellos que determinan el peso o nivel de relevancia de cada variable aleatoria en la correspondiente relación causal difusa. Con el objetivo de reducir parte de la subjetividad del proceso, concretamente la relativa a la determinación de la distribución estadística de los pesos de las variables, dichos pesos son estimados mediante un Algoritmo Evolutivo Multi-objetivo, siempre y cuando existan datos reales disponibles sobre la variable identificada como output en la relación difusa correspondiente.

Los resultados obtenidos mediante este procedimiento de cálculo son evaluados empleando el índice de Shannon (Shannon, 1948), también denominado Entropía de Shannon. Este índice permite medir la incertidumbre asociada a la función de distribución conjunta P de la red bayesiana, estimada conforme al procedimiento de cálculo diseñado. De esta forma, se desarrolla una teoría T sobre la incertidumbre plenamente operacional según las condiciones establecidas en la Teoría de la Información Generalizada (Klir, 2006).

La operatividad de la teoría T propuesta en esta investigación ha sido probada mediante su aplicación al caso de estudio de los flujos de remesas, fenómeno en el que la falta o escasez de datos constituye una importante fuente de incertidumbre, con implicaciones en la construcción y evaluación de modelos para la estimación de dichos flujos. Concretamente, la teoría T ha sido empleada para la evaluación de los flujos de remesas en dos estudios relativos al corredor España-Ecuador, uno de los principales canales de remesas en el área de América Latina y el Caribe y con uno de los crecimientos más importantes tanto en número de personas como de flujos monetarios remitidos durante el periodo 2000-2010 (Ratha *et al.*, 2010).

En el primero de estos estudios, se ha aproximado el flujo de remesas enviado por la comunidad ecuatoriana residente en España durante el periodo 2000-2010. El conocimiento recogido en la literatura sobre el fenómeno de las remesas y las migraciones ha fundamentado la construcción de una red bayesiana, en la que las relaciones causales han sido caracterizadas mediante los axiomas correspondientes a la Teoría de la Probabilidad y de la Lógica Difusa. La probabilidad conjunta P de dicha red ha sido estimada mediante el procedimiento de cálculo propuesto, determinando la incertidumbre asociada a tal estimación mediante la Entropía de Shannon.

En este estudio, la teoría T sobre la incertidumbre permite explorar y contrastar las diversas teorías económicas que explican el fenómeno de las remesas. La metodología diseñada posibilita la experimentación en aquellos supuestos donde la escasez de datos puede resultar un obstáculo para la aplicación de técnicas convencionales, de forma que se abre una nueva vía para la generación de teorías, con base empírica, sobre aspectos concretos de estos flujos financieros.

Por otra parte, la teoría T también permite la realización de pronósticos sobre los flujos de remesas que pueden ser recibidos en una economía, empleando para ello el modelo resultante, calibrado mediante la aplicación iterativa del procedimiento de cálculo diseñado. En esta misma línea, el modelo resultante puede ser empleado para

el análisis de escenarios sobre la evolución futura de estos flujos, aportando información para la toma de decisiones sobre fenómenos complejos, como el mencionado, donde interaccionan multitud de variables y en los que, además, las relaciones entre estas variables pueden ser de cualquier tipo.

En el segundo de los estudios, se han determinado los niveles de relevancia (*pesos*) de las variables que explican el tamaño de la población migrante ecuatoriana en España al inicio del periodo de estudio. Para ello se ha diseñado y desarrollado un Algoritmo Evolutivo Multi-objetivo que, una vez integrado en la teoría T , permite aproximar la combinación óptima de dichos pesos para la correspondiente sección *d-separada* de la red bayesiana. Este procedimiento es perfectamente generalizable para el análisis de otros fenómenos representados mediante redes bayesianas.

La inclusión del Algoritmo Evolutivo Multi-objetivo en el procedimiento de cálculo de la teoría T permite, además de la ya comentada reducción de la subjetividad del procedimiento, el establecimiento de un orden de prelación entre las variables del modelo, basado en el nivel de relevancia de las mismas. Esta ordenación resulta de gran utilidad para las entidades públicas o privadas, interesadas en la gestión de los flujos de remesas, ya que permite identificar cuáles son las variables con mayor impacto en la determinación de estos flujos y, por tanto, sobre las que deben concentrar sus esfuerzos. De esta forma, resulta posible llevar a cabo una gestión más eficiente de los recursos disponibles.

1.2. Introducción al estudio de los flujos de remesas

El volumen y las características de las remesas internacionales ha convertido a estos flujos en una de las cuestiones claves para el desarrollo económico de los países receptores (Adams, 2009). Las últimas estimaciones disponibles para el año 2012, cifran los flujos de remesas recibidos en los países en desarrollo en 406.000 millones de dólares (Ratha *et al.*, 2012), lo que equivale a más de tres cuartas partes los flujos

de remesas a nivel mundial durante el mismo periodo. Este montante iguala al de los flujos de inversión extranjera directa en los países receptores y supera ampliamente la ayuda oficial al desarrollo que éstos reciben. La importancia de estos flujos no sólo radica en su magnitud, que puede llegar a superar el 20% del PIB de los países receptores (Ratha *et al.*, 2012), sino también en su naturaleza anticíclica (Frankel, 2011) y en su estabilidad a lo largo del tiempo (Buch y Kuckulenz, 2010), hecho que se ha podido constatar durante la reciente crisis económica y financiera (Mohapatra *et al.*, 2010).

Por ello, las remesas se han convertido en una importante fuente de financiación externa que contribuye a la estabilidad macroeconómica de los países receptores (Chami *et al.*, 2012). Esta financiación también promueve la sostenibilidad de su deuda externa y la mejora de las calificaciones de dicha deuda por parte de las agencias de rating (Ratha *et al.*, 2011), facilitando las condiciones de acceso a la financiación en los mercados internacionales (Al-Sakka y ap Gwilym, 2010). Sin embargo, estos flujos también pueden tener efectos negativos sobre las economías receptoras (Chami *et al.*, 2009), tales como el incremento de la inflación (Ball *et al.*, 2013; Ulyses Balderas y Nath, 2008), la apreciación de los tipos de cambio y la consiguiente pérdida de competitividad en los mercados internacionales (Acosta *et al.*, 2009; Bourdet y Falck, 2006) o, incluso, la reducción de la oferta de mano de obra como consecuencia del incremento de los salarios de reserva (Amuedo-Dorantes y Pozo, 2006a; Funkhouser, 2006).

Los efectos de los flujos de remesas son también evidentes a nivel microeconómico, ya que estos no son sino la agregación de los millones de transferencias individuales que los migrantes realizan con destino a familiares o amigos residentes en sus países de origen (Adams, 2009). Las remesas permiten que los hogares receptores incrementen su gasto en bienes de consumo e inversión (Adams y Cuecuecha, 2010), especialmente en educación y salud (Calero *et al.*, 2009; López-Cevallos y Chi, 2012), al tiempo que inciden positivamente en su capacidad de ahorro (Gupta *et al.*, 2009), pudiendo hacer frente a imprevistos y desastres naturales

(Yang y Choi, 2007; Yang, 2008) y generando sistemas informales de protección social (Brown *et al.*, 2013). Además de estos efectos directos e indirectos sobre el bienestar de la población del país receptor, las remesas juegan un papel relevante en la reducción de los niveles de pobreza de las comunidades de origen de los migrantes (Adams y Page, 2005; Gupta *et al.*, 2009), por lo que se las considera como el vínculo más tangible y menos controvertido entre migración y desarrollo (Ratha y Shaw, 2007; Russell, 1986).

La potencialidad de los flujos de remesas como herramienta para el desarrollo queda supeditada al diseño, desarrollo e implementación de políticas orientadas a posibilitar y potenciar los efectos positivos de estos flujos y/o a minimizar o contrarrestar los efectos negativos que pudieran ocasionar. Para el diseño de dichas medidas resulta necesario incorporar estimaciones sobre los flujos de remesas recibidas en una economía y su posible evolución. Estas estimaciones suelen ser realizadas tanto por los bancos centrales de los países emisores y receptores de remesas, para su inclusión en la balanza de pagos, como por los organismos internacionales, como el Fondo Monetario Internacional o el Banco Mundial, empleando para ello distintas metodologías, en función de la información disponible.

Es posible distinguir tres enfoques en la estimación de remesas (FMI, 2009). Un primer enfoque, como el vigente en Venezuela, determina los flujos de remesas a partir de datos primarios sobre transacciones realizadas a través de los bancos comerciales, entidades cambiarias autorizadas o agencias de envío de remesas. Este primer método no resulta adecuado cuando los envíos de remesas se llevan a cabo por canales informales. Un segundo enfoque, vigente en países como Ghana, emplea la información obtenida mediante encuestas realizadas tanto a migrantes que envían dinero como a los hogares receptores de remesas. Esta información sí permite la estimación de los flujos enviados por canales informales, pero presenta el inconveniente de su baja periodicidad, ya que el lapso de tiempo entre dos encuestas puede ser superior al año. Un tercer enfoque estima los flujos de remesas a partir de datos secundarios, con los que se construyen distintos tipos de modelos, entre los que

cabe señalar los modelos econométricos como el empleado por el Banco de España o el diseñado por el Banco Mundial, los modelos demográficos¹ como el usado en Estados Unidos de Norte América y los modelos residuales², como el utilizado por el Banco de Albania. Frente a las desventajas que supone el uso de información secundaria, este tercer enfoque resulta de utilidad cuando la medición directa de las transacciones de remesas resulta imposible o inadecuada o bien cuando la realización de encuestas es costosa o impracticable por razones institucionales o culturales, tales como la evasión de impuestos, la economía sumergida o la inmigración ilegal, que hacen poco probable que los hogares revelen información veraz sobre el proceso de envío de remesas.

Bajo cualquier de los enfoques descritos la estimación de los flujos de remesas está condicionada a la fiabilidad, disponibilidad y accesibilidad de datos sobre los envíos de remesas o sobre las variables determinantes de dichos envíos (FMI, 2009). Sin embargo, existe un vasto cuerpo teórico sobre este fenómeno recogido en la literatura científica sobre remesas y migraciones, así como numerosos informes técnicos, susceptibles de ser utilizados para el diseño y evaluación de modelos causales como el que se propone en esta investigación, ofreciendo así una nuevo enfoque para el estudio de los flujos de remesas.

1.3. Estructura de la investigación

Este trabajo de investigación se estructura en seis epígrafes. Tras esta Introducción (Capítulo I) se exponen las hipótesis y se definen los objetivos generales e

¹ Estos modelos determinan los flujos de remesas a partir de datos sobre el tamaño de la población remitente y sobre la cantidad de dinero enviada por término medio. Estos datos son obtenidos a partir de registros administrativos, censos de población y encuestas a la población inmigrante.

² Los modelos residuales estiman las remesas a partir de datos de entradas y salida de divisas obtenidos de la balanza de pagos, sus correspondientes entidades contables, y la asunción de que toda entrada de divisas se convierte en moneda nacional para poder ser utilizada en el país receptor, de tal forma que los desequilibrios existentes son ocasionados por flujos de remesas no observados.

instrumentales del estudio (Capítulo II). A continuación, en la Metodología (Capítulo III) se describe el marco teórico de esta investigación y se desarrolla la propuesta metodológica híbrida que conforma la teoría T sobre la incertidumbre. Seguidamente se presentan los Resultados (Capítulo IV) de los dos estudios de aplicación de la teoría T propuesta. En el epígrafe Discusión se realiza una valoración de los resultados obtenidos y se detalla la producción científica derivada de esta investigación (Capítulo V). Finalmente, se exponen las Conclusiones y las líneas futuras de desarrollo de este trabajo (Capítulo VI). Todas las referencias bibliográficas del texto quedan recogidas de forma exhaustiva en el epígrafe dedicado a la Bibliografía.

II. HIPÓTESIS Y OBJETIVOS

2.1. Hipótesis

Las hipótesis de partida que se pretenden contrastar en esta investigación son las siguientes:

- 1) Existe un procedimiento híbrido para aproximar el valor de la función de distribución de probabilidad conjunta P asociada a una red bayesiana, en la cual las variables aleatorias son continuas y su estructura desconocida, y las relaciones causales definidas son de naturaleza imprecisa.
- 2) Existe un procedimiento que permite determinar la importancia relativa de las variables explicativas en fenómenos causales como los anteriores.
- 3) La Teoría de la Información Generalizada de Klir (2006) es de aplicación a la modelización de fenómenos caracterizadas por una elevada fragmentación del conocimiento existente y/o escasez de datos.

2.2. Objetivos generales e instrumentales

Teniendo en cuenta la hipótesis de partida, los objetivos generales del trabajo son fundamentalmente tres, siendo desarrollados mediante los objetivos instrumentales correspondientes:

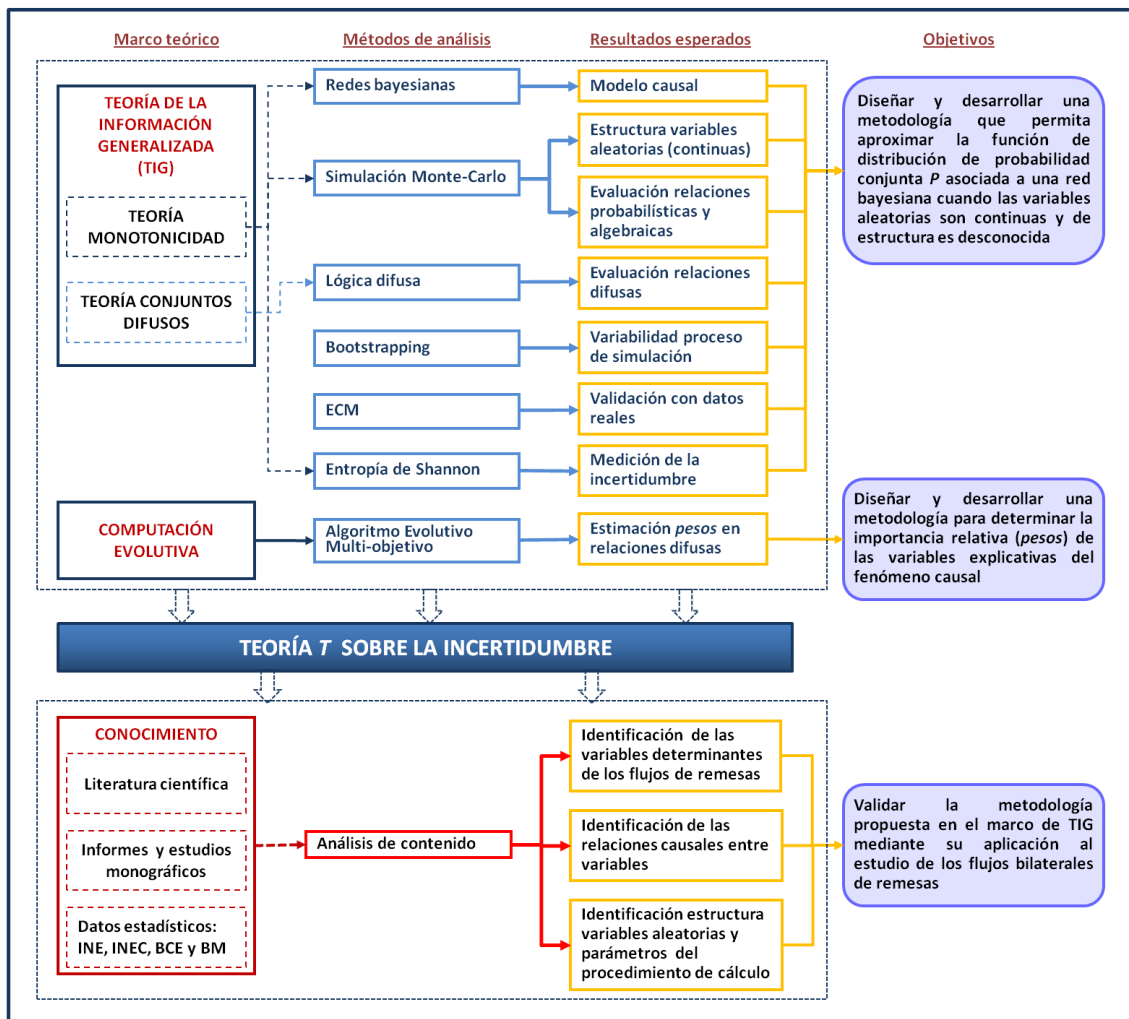
- 1) Diseñar y desarrollar una metodología que permita aproximar la función de distribución de probabilidad conjunta P asociada a una red bayesiana cuando las variables aleatorias son continuas y de estructura es desconocida.
 - 1.1) Identificar las metodologías de cálculo que resultan necesarias para la estimación de P .

- 1.2) Diseñar y desarrollar una metodología híbrida a partir de las metodologías identificadas.
- 2) Diseñar y desarrollar una metodología para determinar la importancia relativa de las variables explicativas del fenómeno causal que dicha red representa.
 - 2.1) Planteamiento del problema multi-objetivo asociado a la determinación de la importancia relativa de las variables.
 - 2.2) Diseño y programación del algoritmo evolutivo multi-objetivo.
- 3) Validar la metodología propuesta en el marco de la Teoría de la Información Generalizada mediante su aplicación al estudio de los flujos bilaterales de remesas.
 - 3.1) Crear un modelo conceptual en forma de red bayesiana
 - 3.2) Identificar la estructura (distribuciones estadísticas) que determina el comportamiento y la evolución temporal de las variables de la red bayesiana.
 - 3.3) Identificar las distribuciones estadísticas que definen la importancia relativa de las variables en la red bayesiana.
 - 3.4) Identificar y estimar los parámetros para el diseño automático de la base de conocimiento (reglas) que permitan la evaluación de las relaciones causales definidas en la red bayesiana.
 - 3.5) Estimar los flujos de remesas.
 - 3.6) Analizar la estabilidad del procedimiento de cálculo.
 - 3.7) Contrastar los resultados obtenidos con los datos recogidos en fuentes secundarias disponibles.

3.8) Calcular la incertidumbre asociada a los resultados obtenidos.

En la Figura 1, se refleja de forma esquemática las principales metodologías y los resultados esperados más relevantes para alcanzar los objetivos generales de esta investigación.

Figura 1. Esquema metodológico de la investigación



Fuente: Elaboración propia.

III. METODOLOGÍA

3.1. La Teoría de la Información Generalizada

La Teoría de la Información Generalizada (TIG) (Klir, 2006) puede ser considerada como un programa de investigación cuyo objetivo fundamental es desarrollar teorías que permitan abordar formalmente la incertidumbre asociada a las situaciones de decisión en cualquier ámbito de la ciencia. En este marco, la concepción probabilística que relaciona incertidumbre y aleatoriedad es sustituida por la relación entre incertidumbre e información.

Así pues, de acuerdo con la TIG la incertidumbre es el resultado de alguna deficiencia en la información disponible sobre el sistema o modelo que conceptualiza la situación de decisión. Esta incertidumbre admite una graduación determinada por la cantidad y calidad de la información disponible sobre dicho sistema (Walker *et al.*, 2003). Esta graduación parte de la situación ideal de determinismo absoluto, en la que se dispone de toda la información o conocimiento sobre el sistema. A continuación aparece la incertidumbre estadística en la que la información sobre el sistema se expresa en términos probabilísticos. El siguiente grado es la incertidumbre de escenarios en las que la información sobre el sistema es expresada de manera vaga o imprecisa. Un grado superior de incertidumbre lo representa la ignorancia reconocida, en la que se carece de información sobre el sistema. En aquellos supuestos en los que el desconocimiento puede ser superado mediante la investigación, proporcionado al menos cierto conocimiento sobre los mecanismos esenciales de la realidad, esta ignorancia es calificada de reducible, en caso contrario es calificada como ignorancia irreducible o indeterminación. El último escalón de la graduación corresponde a la ignorancia total, el grado más profundo de incertidumbre, en el cual ni siquiera es posible determinar qué es lo desconocido.

Para la TIG, la información relativa a una situación de decisión, obtenida mediante una determinada acción (experimentación, recopilación de conocimiento, etc.), es medida como la diferencia entre la incertidumbre a priori y la incertidumbre a

posteriori que existe en esa situación. Esta operación da lugar al concepto de información basado en la incertidumbre (Klir, 2006). La utilidad de medir la información generada radica en la posibilidad de comparar la capacidad informativa de los distintos sistemas o modelos desarrollados para un determinado problema.

Por otra parte, las distintas deficiencias de información dan lugar a distintos tipos de incertidumbre, que incluso pueden producirse de forma simultánea (Celikyilmaz y Türksen, 2009). Así pues, resulta posible distinguir entre la incertidumbre asociada a las limitaciones o errores de medición, la incertidumbre provocada por la ambigüedad o vaguedad consustancial a procesos cognitivos expresados en lenguaje natural, la incertidumbre como consecuencia la existencia de evidencias contradictorias o la incertidumbre debida a la complejidad del fenómeno objeto de estudio (Zimmermann, 2000).

Como consecuencia de la existencia de distintos tipos de incertidumbre y de las correspondientes teorías para su representación formal (Dubois y Prade, 2010), la TIG propugna que la elección de la correspondiente teoría debe estar guiada por la naturaleza del problema a resolver, permitiendo, en cualquier caso, expresar la ignorancia completa sobre el sistema, así como no ignorar ninguna de la información disponible sobre el mismo (Klir, 2006). Para ello la TIG se basa en dos importantes generalizaciones acontecidas en el campo de las matemáticas durante la segunda mitad del siglo XX. Por una parte, la generalización de la Teoría de Medida (Halmos, 1950) a la Teoría de la Monotonidad, propuesta por primera vez por Choquet (1953) y, por otra, la generalización de la Teoría de Conjuntos a la Teoría de Conjuntos Difusos (Zadeh, 1965).

El desarrollo de teorías T sobre la incertidumbre en el marco de la TIG exige de la consecución de una serie de condiciones, garantizando que las teorías desarrolladas sean capaces de hacer frente a los distintos tipos de incertidumbre identificados en las situaciones decisionales. Así pues, una teoría T sobre la incertidumbre es plenamente

operacional cuando cumple las siguiente cuatro premisas (Klir, 2006):

- Permite la selección del lenguaje formal apropiado para expresar los distintos tipos de incertidumbre, de tal forma que las funciones inciertas, denotadas por u , pueden ser caracterizadas mediante los axiomas apropiados.
- Desarrolla una metodología que permite afrontar las distintas situaciones o problemas en los que la teoría de la incertidumbre desarrollada resulta de aplicación.
- Desarrolla de un procedimiento de cálculo que permite un tratamiento adecuado de las funciones u .
- Desarrolla una medida funcional de la incertidumbre U asociada a cada u de la correspondiente teoría.

Como se indicó en el apartado introductorio, este trabajo se enmarca dentro del programa de investigación de la TIG. Por ello, la exposición de la propuesta metodológica a lo largo de los siguientes apartados se ordena y desarrolla según las condiciones señaladas anteriormente.

3.2. Axiomas para la formalización de la incertidumbre en la teoría T

En las redes bayesianas, el modelo conceptual relativo al fenómeno objeto de estudio se conforma a partir de las variables explicativas del dominio de estudio y de las relaciones de influencia o causalidad existentes entre dichas variables. La Teoría de la Probabilidad se utiliza para expresar, siempre que sea posible, la incertidumbre sobre el fenómeno conceptualizado mediante dicho modelo, convirtiéndolo así en una red bayesiana. Las probabilidades marginales de las variables aleatorias del modelo expresan la incertidumbre sobre los posibles valores de estas variables, mientras que las probabilidades condicionadas expresan la incertidumbre sobre las relaciones causales identificadas. En el caso de redes bayesianas construidas a partir del

conocimiento experto, sobre las que versa esta investigación, tanto el modelo causal como las probabilidades son asignadas por los expertos del dominio o extraídas de la literatura (Druzdzel y van der Gaag, 2000) empleando para ello diversas técnicas (Hu et al., 2007; Nadkarni y Shenoy, 2004).

3.2.1. Determinación de la estructura de las variables aleatorias de la red bayesiana

El conocimiento disponible permite identificar las variables explicativas del fenómeno objeto de estudio. La naturaleza estocástica de estas variables obliga a su caracterización mediante la correspondiente distribución de probabilidad, expresando así la incertidumbre probabilística sobre los distintos valores que dicha variable puede adoptar.

La estructura de estas variables puede ser determinada a partir de distintas fuentes. Por un lado, estas distribuciones podrían venir identificadas en la literatura científica o en los informes técnicos sobre el dominio de estudio (Druzdzel y van der Gaag, 2000). Por otro, si la ausencia de datos sobre todas o algunas de las variables del modelo no es total, pueden emplearse algunas técnicas estándar para la determinación de la distribución teórica que mejor ajusta dichos datos, como el método de máxima verosimilitud, el método de los mínimos cuadrados, el método de los momentos o el método de los percentiles (Biller y Gunes, 2010).

Finalmente, si los datos son muy escasos, se desconfía de su fiabilidad o se trata de datos históricos que no reflejan bien la situación actual, el histograma de datos disponibles puede ser utilizado para aproximar las distribuciones estadísticas correspondientes (Kelton, 2009), normalmente de tipo no paramétrico.

3.2.2. Definición de las relaciones causales de la red bayesiana

Las relaciones causales entre las variables de la red bayesiana también son identificadas a partir del conocimiento, tanto el elicitado de los expertos del dominio, como el proveniente de la literatura científica, donde estas relaciones son establecidas mediante otros modelos sobre el fenómeno o sobre modelos parciales sobre algunas características de dicho fenómeno. En este último caso, nos encontramos ante un conocimiento fragmentado, en el que las relaciones pueden ser de tipo probabilístico, pueden venir definidas por una forma funcional concreta, o pueden ser expresadas en lenguaje natural. En este punto, aparece otro tipo de incertidumbre en el sistema, motivado por la imprecisión o vaguedad del lenguaje, en lugar de por la aleatoriedad característica de las relaciones probabilísticas. Este nuevo tipo de incertidumbre es abordado, bajo el paradigma de la TIG, mediante la lógica difusa.

De esta forma, las relaciones causales del modelo son clasificadas en tres grupos, aquellas que son expresadas mediante la probabilidad condicionada y que reciben el nombre de probabilísticas, aquellas que poseen una forma funcional definida y que denominamos algebraicas, y aquellas cuya existencia es declarada mediante el lenguaje natural, de una forma imprecisa, y que denominamos difusas.

Las relaciones de tipo probabilístico son la forma habitual de modelar las relaciones entre variables de una red bayesiana. En las redes bayesianas construidas en base a conocimiento, las probabilidades condicionadas de cada variable de la red aleatoria son determinadas por el experto, reflejando el grado de creencia sobre los posibles valores de dicha variable a la vista de los estados de aquellas variables identificadas como causas de la primera (Chen y Pollino, 2012). Es muy frecuente que, en la determinación de estas probabilidades, las variables aleatorias sean de naturaleza discreta o sean discretizadas mediante métodos convencionales, con la consiguiente pérdida de información (Uusitalo, 2007).

Un ejemplo de relación algebraica viene recogido en Straubhaar (1986) “...Mediante la combinación de los determinantes de (i) la cantidad individual de ingreso de la que dispone un individuo para remitir [$s(Y)$], (ii) la decisión individual de remitir parte del dinero [$r(e,i)$], y (iii) el número total de emigrantes que potencialmente podrían remitir [$M(E)$], el flujo total de remesas [(R)] hacia el país de emigración viene dado por $R = s(Y) \times r(e,i) \times M(E)$ ”. Estas relaciones algebraicas, denotadas como $Y = f(X_i)$, resultan fácilmente evaluables mediante modelos computacionales basados, por ejemplo, en la simulación Monte Carlo, tanto si las variables aleatorias son de naturaleza discreta o continua.

Una relación difusa puede ser ilustrada mediante, por ejemplo, la declaración causal que aparece en Freund y Spatafora (2008) “...los costes de envío de remesas son menores a medida que los sistemas financieros están más desarrollados y los tipos de cambio son menos volátiles...”. Estas relaciones difusas, se denotan como $Y = u(X_i)$, y son modeladas mediante lógica difusa.

3.2.2.1. Modelado difuso de relaciones causales en redes bayesianas

Mediante la Lógica difusa, las relaciones causales difusas $Y = u(X_i)$ son descritas en términos semánticos, empleando para ello la estructura condicional de las reglas difusas estándar del tipo “IF-THEN”:

$$\text{IF } x_i \in X_i \text{ isr } A_i \text{ THEN } y \in Y \text{ isr } B \quad (1)$$

Una regla difusa no es más que una proposición difusa múltiple, formada mediante la conexión de proposiciones difusas atómicas. Estas proposiciones atómicas son definidas sobre los distintos valores de las variables que intervienen en la relación. La proposición de la izquierda se denomina *antecedente* o *premisa*, en nuestro caso $x_i \in X_i \text{ isr } A_i$, mientras que la de la derecha se conoce como *consecuente* o *conclusión*,

$y \in Y$ is $r B$. Las relaciones entre estas proposiciones se establecen mediante la aplicación del conectivo lógico de implicación (\rightarrow), cuyos análogo natural es *IF-THEN*.

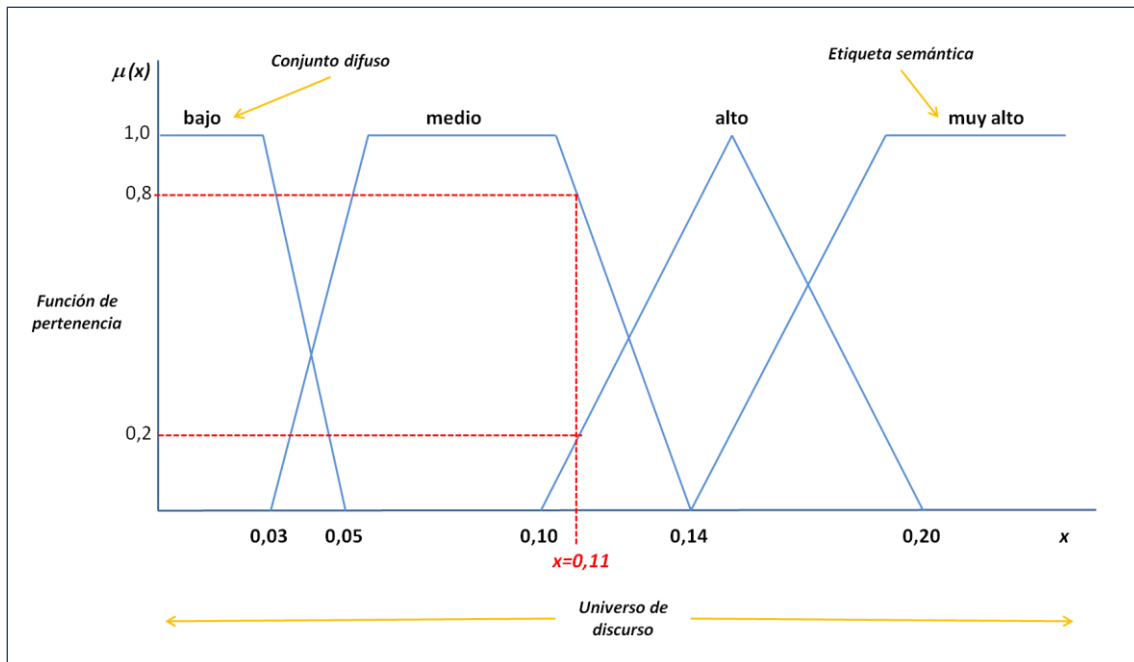
Las proposiciones difusas atómicas o simples permiten expresar el valor numérico de las variables en etiquetas semánticas que, en el ámbito de la lógica difusa, reciben el nombre de conjuntos difusos (Zadeh, 1965). Así pues, en la proposición difusa atómica $x_i \in X_i$ is $r A_i$, la variable correspondiente toma valores x_i en el universo de discurso X_i , el cual puede ser descrito mediante diferentes etiquetas lingüísticas o conjuntos difusos. Cada uno de estos valores lingüísticos recibe el nombre de conjunto difuso A_i . El conjunto difuso se define mediante una función que asocia cada elemento del universo de discurso su grado de pertenencia $\mu_{A_i}(x_i)$ a dicho conjunto, en el intervalo $[0,1]$, de tal forma que el conjunto difuso puede ser denotado como $\mu_{A_i}(x_i)$.

A modo de ejemplo, podríamos considerar una de las variables determinantes de los flujos de remesas como es el “coste de envío”, expresada en tanto por uno y cuyo universo de discurso viene definido por el intervalo $[0, 0,25]$. Sobre este universo de discurso se han definido cuatro conjuntos difusos, “bajo”, “medio”, “alto” y “muy alto”, tal y como se muestra en la Figura 2. Ante un valor concreto de coste de envío, por ejemplo 0,11, se podría decir que el coste de envío es medio, con un grado de pertenencia de 0,8, pero también que dicho coste es alto con un grado de pertenencia 0,2.

La función de pertenencia asociada un conjunto difuso puede adoptar distintas formas, siendo las más comunes las triangulares, trapezoidales, gaussianas y sigmoidales. Éstas últimas se suelen usar para calificar valores lingüísticos extremos, tales como bajo y muy alto en la Figura 2, mientras que el resto se suelen usar para describir valores intermedios, como medio (trapezoidal) y alto (triangular), en la Figura 2. A diferencia de las funciones triangulares, el uso de la funcione trapezoidales implica un margen de tolerancia alrededor del valor que se considera como más representativo del valor lingüístico asociado al conjunto difuso. En la Figura 2 los

conjuntos difusos aparecen solapados dos a dos, de forma que la suma de las funciones de pertenencia de ambos conjuntos tiene como resultado la unidad, si bien el solapamiento puede darse entre un mayor número de conjuntos difusos y la suma de las correspondientes funciones de pertenencia puede ser distinta a la unidad.

Figura 2. Conjuntos difusos definidos sobre la variable aleatoria “coste de envío”



Fuente: Elaboración propia.

Por tanto, si A_i es un conjunto difuso, la verdad de una proposición como " $x_i \in X_i$ isr A_i " o "0,11 es un coste de envío medio", determina en la lógica tradicional un grado de verdad absoluto $\tau(A_i(x)) = \tau \in \{0,1\}$, mientras que en la lógica difusa el grado de verdad sería $\tau(A_i(x)) = \tau \in [0,1]$. Mediante los grados de verdad de una proposición se incorpora la incertidumbre al razonamiento, lo que permite el tratamiento de conceptos imprecisos expresados en lenguaje natural.

Las reglas difusas "IF-THEN", que definen las relaciones causales difusas de la red bayesiana, permiten describir formalmente la incertidumbre causada por la definición vaga o imprecisa de las relaciones causales del tipo "si el coste en envío es alto las remesas son bajas". Así pues, una regla difusa representa una relación difusa

entre el antecedente y el consecuente, cuya función de pertenencia viene dada por la siguiente expresión, donde Φ representa el operador de implicación:

$$\mu_{A_i \rightarrow B}(x_i, y) = \Phi(\mu_{A_i}(x_i), \mu_B(y)), \forall x_i \in X_i, \forall y \in Y \quad (2)$$

La generalización de los métodos de inferencia de la lógica proposicional clásica³, permite aplicar los procesos de inferencia a las reglas difusas. Zadeh (1975) propuso obtener la conclusión del *modus ponens generalizado* mediante la siguiente expresión, conocida como *regla de inferencia composicional*:

$$B' = A'_i \circ (A \rightarrow B) \quad (3)$$

Esta regla de inferencia composicional describe un mecanismo de inferencia que permite obtener la conclusión o consecuente de una regla difusa, pudiendo ser generalizado para un conjunto de reglas. La función de pertenencia asociada a la formulación anterior, cuyo resultado es el grado de verdad de la regla difusa $\tau \in [0,1]$, se detalla a continuación:

$$\mu_{B'}(y) = \sup_{x_i \in X_i} \text{AND}(\mu_{A'_i}(x_i), \mu_{A_i \rightarrow B}(x_i, y)), \forall x_i \in X_i, \forall y \in Y \quad (4)$$

donde *sup* representa al operador *max* y *AND* representa al operador *min*, por lo que este tipo de inferencia recibe el nombre de inferencia *max-min*.

³ El *modus ponens generalizado* establece que, dada la regla difusa « *IF* $x_i \in X_i$ *ISR* A_i *THEN* $y \in Y$ *ISR* B » y la observación « $x_i \in X_i$ *ISR* A'_i », se puede obtener la conclusión « $y \in Y$ *ISR* B' », donde el conjunto difuso B' es más próximo a B cuanto más próximo sea A'_i al conjunto A_i . Por su parte, el método *modus tollens generalizado* establece que, dada la regla difusa « *IF* $x_i \in X_i$ *ISR* A_i *THEN* $y \in Y$ *ISR* B » y la observación « $y \in Y$ *ISR* B' », se puede obtener la conclusión « $x_i \in X_i$ *ISR* A'_i », donde el conjunto difuso A'_i es más diferente a A_i cuanto más diferente sea B' al conjunto B . La utilización de *modus ponens* para obtener conclusiones se conoce como inferencia hacia adelante (*forward inference*) mientras que el uso de *modus tollens* se denomina inferencia hacia atrás (*backward inference*).

Finalmente, cabe señalar que cada relación causal difusa del modelo no viene definida por una única regla difusa, sino por un conjunto de reglas R . Este conjunto de reglas conforma la base de conocimiento a priori sobre la que opera el motor de inferencia difusa, recogiendo, a ser posible, todas las posibles combinaciones de valores de las variables inputs y un único output (Zadeh, 1965):

$$R: \text{ALSO} \left[\text{IF } \text{AND}_{i=1}^n (x_i \in X_i \text{ isr } A_{id}) \text{ THEN } y_d \in Y_d \text{ isr } B_d \right], d=1,2,\dots,c \quad (5)$$

donde $d=1,2,\dots,c$ es el número de reglas que define la relación causal difusa, $i=1,2,\dots,n$ el número de inputs de la relación difusa, *isr* el conector semántico, A_{id} la función de pertenencia para las variables inputs X_i en cada regla difusa, *AND* el conector que agrega los valores de pertenencia de los inputs en la correspondiente regla, B_d la función de pertenencia para el output Y , ya que cada regla d tiene un único output y_d , *ALSO* el conector que agrega los output de cada una de las reglas en R . En la Tabla 1, se recoge algunos ejemplos de este tipo de reglas de base semántica en el que intervienen operadores difusos.

Tabla 1. Conjunto de reglas que definen una relación causal difusa

Relación	$Y=u(X_i, X_j)$		
Variable	X_i	X_j	Y
Conjuntos difusos	ELEVADO	ALTO	ALTO
	MEDIO	MEDIO	MEDIO
	BAJO	BAJO	BAJO
<i>Algunos ejemplo de reglas difusas IF...THEN para la relación causal difusa $Y=u(X_i, X_j)$</i>			
<i>IF X_i isr ELEVADO AND X_j isr ALTO THEN Y isr MEDIO</i>			
<i>IF X_i isr ELEVADO AND X_j isr BAJO THEN Y isr ALTO</i>			
<i>IF X_i isr MEDIO AND X_j isr BAJO THEN Y isr MEDIO</i>			
<i>IF X_i isr MEDIO AND X_j isr MEDIO THEN Y isr MEDIO</i>			
<i>IF X_i isr BAJO AND X_j isr ELEVADO THEN Y isr BAJO</i>			
<i>IF X_i isr BAJO AND X_j isr MEDIO THEN Y isr BAJO</i>			

Fuente: Elaboración propia.

3.2.3. Identificación de las variables y sus relaciones causales en el caso de los flujos de remesas

La identificación de las variables y las relaciones causales, que dan lugar a la red bayesiana relativa a los flujos de remesas, se ha realizado en base al conocimiento recogido en el vasto, aunque fragmentado, cuerpo teórico sobre los fenómenos de remesas y migraciones, puesto que cualquier definición relativa a los flujos de remesas se construye sobre estos dos elementos.

La revisión de la literatura económica sobre el fenómeno de las remesas ha permitido identificar aquellas variables que determinan el envío de remesas a los países de origen, tanto a nivel macroeconómico como microeconómico, así como las relaciones causales entre las mismas. Aunque se han desarrollado distintas teorías para explicar los flujos de remesas (De Haas, 2010), esta investigación se ha nutrido de los estudios desarrollados bajo el enfoque de la Nueva Economía sobre las Migraciones Laborales (Lucas y Stark, 1985; Taylor, 1999), ya que muestran mayor desarrollo tanto a nivel teórico (Rapoport y Docquier, 2006) como empírico (Carling, 2008; Hagen-Zanker y Siegel, 2007) sobre los distintos determinantes de estos flujos.

Los estudios que abordan el fenómeno de la migración, desde la óptica de los flujos internacionales de personas, han permitido incorporar al modelo conceptual aquellos factores determinantes del tamaño de la población migrante en el país de acogida y sus relaciones. Al igual que en el caso anterior, también existen distintas teorías que explican las causas de los flujos migratorios (De Haas, 2010; Massey *et al.*, 2005). Sin embargo, en esta ocasión, el modelo ha sido construido combinando factores explicativos de distintas teorías, tales como la Teoría Neoclásica (Borjas, 1989; Harris y Todaro, 1970; Todaro, 1969), la Teoría de Mercado Laboral Dual (Piore, 1979) y la Nueva Economía de las Migraciones Laborales (Stark y Bloom, 1985). Por otra parte, aquellos estudios que abordan el fenómeno de la migración desde la óptica del mercado laboral, especialmente aquellos relativos a la asimilación salarial de los

migrantes en los países de acogida (Borjas, 2001; Chiswick, 1978), han sido utilizados para identificar los factores determinantes de los salarios obtenidos por estos migrantes.

En base a los mencionados estudios, los flujos de remesas recibidos en un país Rm_t durante un periodo de tiempo t son definidos como la suma de las transferencias de ingresos realizadas por su población emigrante (Adams, 2009). A partir de esta definición, las remesas pueden ser modeladas mediante la siguiente expresión, dada por Straubhaar (1986), que determina la primera relación causal, de carácter algebraico y denotada como f :

$$Rm_t = f(RPm_t, WAm_t, Nm_t) = RPm_t \times WAm_t \times Nm_t \quad (6)$$

siendo RPm_t la propensión media a remitir del migrante, expresada como un porcentaje sobre el salario medio percibido en el país de destino WAm_t , considerado en la moneda de curso legal del país receptor de los flujos de remesas, y Nm_t el número de nacionales del país receptor que se encuentran en el país emisor de dichos flujos.

A su vez, la propensión media a remitir RPm_t queda definida mediante la siguiente relación difusa, denotada por u :

$$RPm_t = u(FMN_t, TC_t, IRD_t, SPm_t) \quad (7)$$

donde FMN_t representa las necesidades económica de la familia del migrante que permanece en el país de origen; TC_t son los costes de transacción asociados al envío de remesas al país de origen; IRD_t es el diferencial de tipo de interés entre el país de origen y el país donde se encuentra el migrante; y SPm_t es el perfil socio-demográfico

del migrante. Algunas de estas variables son utilizadas como variable proxy de las distintas motivaciones que subyacen al envío de remesas por el migrante a su país de origen.

La situación económica de la familia del migrante en el país de origen, FMN_t , es una de las principales motivaciones para el envío de remesas, revelando un comportamiento altruista por parte del migrante. De esta forma se establece una relación causal, de carácter positivo entre ambas variables, en la que la propensión del migrante a enviar remesas aumenta a medida que lo hacen las necesidades de su familia en el país de origen (Lucas y Stark, 1985; Osili, 2007).

Por su parte, los costes de transacción TC_t se relacionan con el envío de remesas por un motivo de inversión. Así, el migrante es más proclive a enviar remesas cuando se producen determinadas condiciones financieras que le resultan favorables para el envío de remesas, es decir, cuando los costes de transacción son menores. De forma análoga, los estudios empíricos muestran que, cuando estos costes aumentan, los migrantes reducen la frecuencia de sus envíos o los posponen (Hysenbegasi y Pozo, 2006), además de aumentar el uso de canales informales, más económicos pero también más arriesgados (Freund y Spatafora, 2008). La evolución de los costes de transacción depende de las comisiones cobradas por los intermediarios y del tipo de cambio entre los países del corredor. En este sentido, la depreciación de la moneda del país de origen del migrante puede ocasionar un aumento de las cantidades enviadas (Higgins *et al.*, 2004). Sin embargo, los efectos de este segundo factor sobre los flujos de remesas no son concluyentes, ya que otros estudios señalan disminuciones en las cantidades remitidas (Lianos, 1997; Lueth y Ruiz-Arranz, 2007; Straubhaar, 1986) ante depreciaciones de la moneda del país de origen del migrante.

El diferencial de tipo de interés IRD_t , también es una variable indicativa de la motivación inversora del migrante en el envío de remesas (Amuedo-Dorantes y Pozo, 2013; Aydas *et al.*, 2005). En relación a esta variable, la evidencia empírica tampoco

resulta concluyente, puesto que existen corredores en los que los efectos de IRD_t sobre la propensión a remitir son muy débiles o prácticamente inexistentes (Elbadawi and Rocha, 1992; Straubhaar, 1986), mientras que en otros esta relación es relevante y presenta una forma no lineal (Aydas *et al.*, 2005; Higgins *et al.*, 2004). En este último caso, los estudios señalan que un diferencial positivo atrae flujos hacia el país de origen del migrante, pero que un diferencial excesivamente alto puede provocar un descenso en el envío de remesas, al ser interpretado por los migrantes como un signo de inestabilidad económica o institucional en el país

Por su parte, el perfil socio-demográfico del migrante SPm_t , se define a partir de un conjunto de características que la literatura señala como determinantes en el envío de remesas, tales como la edad del migrante Am_t , su nivel educativo Em_t y la duración de la estancia en el país de acogida SDm_t , definiéndose así una nueva relación difusa:

$$SPm_t = u(Am_t, Em_t, SDm_t) \quad (8)$$

Existe una gran cantidad de literatura sobre el efecto de estas variables sobre los flujos de remesas, sin embargo la evidencia empírica resulta contradictoria en la mayoría de los casos. En relación a la edad del migrante, algunos estudios empíricos detectan una relación positiva (Makina, 2012; Merkle y Zimmermann, 1992), mientras que otros apuntan hacia la existencia de una relación en forma de V invertida (Durand *et al.*, 1996; Holst *et al.*, 2012). Igual ocurre con el nivel educativo del migrante, respecto al que se detecta tanto una relación positiva (Bollard *et al.*, 2011; Gibson y McKenzie, 2012), fundamentada en la obtención de mayores ingresos a mayores niveles de instrucción y, por ende, mayor capacidad remesadora, como una relación negativa (Faini, 2007; Niimi *et al.*, 2010), en la que los migrantes con mayores niveles educativos se verían menos apremiados a enviar remesas a sus familiares, dado que suelen proceder de familias acomodadas. La influencia de duración de la estancia en el país de acogida también arroja resultados contradictorios. Así pues, se han detectado

tanto relaciones positivas, asociadas a la intención de reforzar vínculos familiares (Osili, 2007) o generar un clima social favorable de cara a un posible regreso del migrante (Dustmann y Mestres, 2010), como negativas, relacionadas con el debilitamiento de los vínculos con los familiares en el país de origen (Grigorian y Melkonyan, 2011) y la reunificación familiar (Merkle y Zimmermann, 1992). Este segundo grupo de estudios sustenta la hipótesis sobre la reducción de los flujos de remesas por el transcurso del tiempo en el país de acogida (*decay hypothesis*), que ha sido descartada por otros autores (Brown, 1997; Makina, 2012)

Volviendo a las variables que define los flujos de remesas, el salario del migrante WAm_t ha sido destacado por la literatura como otra de las de variables determinantes de los flujos de remesas, apuntando siempre hacia una relación positiva entre ambas variables, (Lueth y Ruiz-Arranz, 2007; Osili, 2007; Rapoport y Docquier, 2006). De cara a estimar los flujos de remesas recibidos en una economía, esta variable se expresa en la moneda de curso legal de la economía receptora de los flujos, de tal forma este salario se calcula mediante una simple multiplicación del salario medio percibido por el migrante WDE_t , expresado en la moneda de curso legal de país de acogida, y el tipo de cambio establecido entre las monedas de ambos países ER_t :

$$WAm_t = f(WDE_t, ER_t) = WDE_t \times ER_t \quad (9)$$

El salario WDE_t que recibe el migrante en el país de destino depende tanto de la situación del mercado laboral de dicho país, aproximada a través de su tasa de desempleo UD_t (Harris y Todaro, 1970), como de su perfil laboral LPM_t (Mincer, 1958). En el primer caso existe una relación negativa entre los salarios de los individuos asentados en un territorio y la tasa de desempleo (Blanchflower y Oswald, 1990), siendo ésta relación más acusada en los supuestos de población migrante y especialmente en presencia de shock económicos (Barth *et al.*, 2004).

$$WDE_t = u(UD_t, LPM_t) \quad (10)$$

El perfil laboral del migrante LPM_t suele tener un efecto positivo sobre el salario del migrante, mediante las distintas variables que lo definen:

$$LPM_t = u(AM_t, EM_t, SDM_t) \quad (11)$$

Según Chiswick (Chiswick *et al.*, 1997), el perfil laboral del migrante puede ser modelado considerando la edad del migrante AM_t , su nivel educativo EM_t , y la duración de la estancia en el país de acogida SDM_t . La edad del migrante AM_t permite aproximar su experiencia laboral (Amuedo-Dorantes y De La Rica, 2007), que es valorada por el empleador, aunque en menor medida que la adquirida en el país de acogida (Chiswick y Miller, 2009). En cuanto al nivel educativo EM_t , los migrantes con mayores niveles educativos presentan mayores tasas de asimilación (Borjas, 2000) a la situación de los trabajadores nativos, en términos de salario (Izquierdo *et al.*, 2009) y de tasas de empleo (Fernández y Ortega, 2008). Por último, la duración de la estancia en el país de acogida SDM_t , permite tanto un mayor conocimiento el mercado laboral como la adquisición de habilidades y experiencia en el país de destino, lo que repercute positivamente en el salario del migrante (Chiswick y Miller, 2009).

La población o stock de migrantes en el país emisor de los flujos de remesas, Nm_t (Freund y Spatafora, 2008; Ratha y Mohapatra, 2010), es la tercera y última de las variables involucradas en la estimación de tales flujos. El stock de población migrante puede ser calculado a partir de la propensión a emigrar MP_t que presenta la población PO_t del país receptor de remesas:

$$Nm_t = f(PO_t, MP_t) = PO_t \times MP_t \quad (12)$$

La propensión a emigrar MP_t se ve influenciada por distintos factores entre los que cabe señalar los beneficios derivados del modelo de bienestar social del país elegido como destino migratorio (Borjas, 1999), el diferencial de ingresos entre este país y el de origen DW_t (Todaro, 1969) y los costes asociados a la migración MC_t (Clark *et al.*, 2004).

$$MP_t = u(SB_t, DW_t, MC_t) \quad (13)$$

Mientras que el diferencial de salarios DW_t ejerce una influencia positiva sobre la propensión a emigrar (Harris y Todaro, 1970), la relación no es tan evidente para los beneficios sociales SB_t , que pueden ser considerados como una forma de ingreso en especie. Mientras que Borjas (1999) señala que las personas suelen emigrar hacia aquellos destinos con mejores programas sociales (*welfare magnet hypothesis*), algunos estudios establecen que su influencia es relativamente baja en comparación con la ejercida por las condiciones del mercado laboral (De Giorgi y Pellizzari, 2009) y otros apuntan hacia la inexistencia de una relación significativa entre los beneficios sociales del país de acogida y los flujos migratorios (Pedersen *et al.*, 2008).

Por otra parte, los costes asociados a la migración MC_t tienden a reducir la propensión de la población a emigrar, ya que junto con los gastos de asociados al desplazamiento, que normalmente aumentan con la distancia (Clark *et al.*, 2004; Pedersen *et al.*, 2008), han de considerarse otros, como por ejemplo los costes administrativos derivados de las políticas migratorias existentes en el país de destino (Bertoli y Fernández-Huertas Moraga, 2012) o los recursos necesarios para el sostenimiento del migrante hasta su incorporación al mercado laboral (Chiswick, 1999). De esta forma, la pobreza del potencial migrante puede constituir una barrera a la decisión migratoria (Clark *et al.*, 2004). Ante esta situación, las redes sociales de migrantes de una misma nacionalidad desempeñan una doble función. Por una parte, pueden reducir las restricciones presupuestarias a las que se enfrenta el migrante mediante la concesión de préstamos (Mckenzie y Rapoport, 2007); y, por otra, pueden

ayudar a reducir los costes monetarios de la emigración y también los no monetarios, derivados de la adaptación del individuo a una nueva cultura, lengua o costumbres (Epstein, 2008).

3.3. Las redes bayesianas como metodología para la representación de la incertidumbre en la teoría T

3.3.1. Elementos de una red bayesiana

Las redes bayesianas son un tipo de modelos que permite representar el conocimiento sobre la realidad objeto de estudio de una forma económica y compacta. Para ello, se sirven de un modelo gráfico, que incluye las variables aleatorias y sus relaciones causales, así como de una distribución de probabilidad conjunta asociada, que recoge la incertidumbre del fenómeno representado (Pearl, 2009).

El modelo gráfico asociado potencialmente a una red bayesiana es un grafo G , donde el conjunto de sus nodos V representa las variables aleatorias del dominio de estudio $X = \{X_1, X_2, \dots, X_n\}$, y el conjunto de conexiones entre estas variables E representa las relaciones de causalidad existentes (Kiiveri *et al.*, 1984; Wermuth y Lauritzen, 1983). El grafo G es un grafo acíclico directo (DAG, por sus siglas en inglés), cuando las conexiones se denotan mediante flechas unidireccionales (grafo directo), de tal forma que las conexiones establecidas entre cada par de variables, representadas como $X_i \rightarrow X_j$, expresan una influencia o relación causal en la que X_i es causa y X_j es el efecto producido. Además en los DAG no se admiten los bucles, esto es, la conexión de un nodo consigo mismo, denotado como $X_i \rightarrow X_i$; ni tampoco la existencia de ciclos (grafo acíclico), representado como $X_i \leftrightarrow X_j$, que expresan relaciones de causalidad mutua o procesos de retroalimentación.

En los DAG, al igual que en cualquier tipo de grafo, es posible trazar caminos o sucesiones ininterrumpidas de conexiones entre nodos, cualquiera que sea el sentido de las flechas, de tal forma que estos caminos identifican una serie de conjuntos específicos de nodos. Así pues, en la conexión $X_i \rightarrow X_j$, se dice que X_i es padre de X_j y, a su vez, que X_j es hijo de X_i . Los antecedentes del nodo X_i , denotados como $ant(X_i)$, son el conjunto de nodos que tienen un camino hasta X_i , mientras que sus descendientes, $des(X_i)$, son el conjunto de nodos para los que X_i es el nodo inicial del camino. En un DAG un nodo se denomina *raíz* si carece de padres y *sumidero* si carece de hijos, todo DAG tiene al menos una *raíz* y un sumidero (Pearl, 2009).

El conjunto de nodos más relevante en una red bayesiana, y en particular en un DAG, es el conocido como Padres Markovianos (*PA*). La relevancia de los *PA* se debe a que permiten simplificar el cálculo de la probabilidad conjunta P de una red bayesiana mediante la aplicación del principio de descomposición. Dicho principio permite descomponer la distribución de probabilidad P definida por n variables discretas X_1, X_2, \dots, X_n en una multiplicación de n distribuciones condicionadas de la forma $P(x_1, x_2, \dots, x_{i-1}) = \prod_i P(x_i | x_1, x_2, \dots, x_{i-1})$.

Así pues, dado el conjunto ordenado de variables $V = \{V_1, V_2, \dots, V_n\}$ en el DAG G y la distribución de probabilidad conjunta de estas variables $P(V)$, se denomina Padres Markovianos de X_i (PA_i) al conjunto mínimo de antecedentes de X_i que hace que X_i sea independiente de cualquier otra variable antecedente, es decir, $P(x_i | pa_i) = P(x_i | x_1, x_2, \dots, x_{i-1})$, de tal forma que la distribución de probabilidad conjunta P del grafo G se puede calcular mediante:

$$P(x_1, x_2, \dots, x_n) = \prod_i P(x_i | pa_i) \quad (14)$$

3.3.2. Redes bayesianas y propiedades de Markov

La condición necesaria y suficiente para que el grafo G sea una red bayesiana es que la distribución de probabilidad conjunta P pueda ser factorizada de acuerdo con la Ec. 14, se dice entonces que G representa a P o que G es compatible con P (Lauritzen, 1996) Esta condición, conocida como Compatibilidad de Markov, implica que el DAG G describe el proceso estocástico real que genera P (Pearl, 2009). La forma de caracterizar el conjunto de distribuciones compatibles con un DAG G consiste en ordenar la secuencia de independencias condicionales que dicho DAG debe satisfacer.

Estas independencias condicionadas pueden ser determinadas a partir del DAG mediante la aplicación del criterio gráfico de separación direccional o *d-separación* (Pearl, 2009, 1988). Así pues, dados tres conjuntos de variables disjuntos X , Y y Z que representan los nodos de un DAG G , X es independiente de Y , dado Z , en cualquier distribución compatible con G , si los nodos correspondientes a Z bloquean (interrumpen el flujo de información o dependencia) todos los caminos desde los nodos de X a los de Y . Un camino p está bloqueado o *d-separado* por un conjunto de nodos Z si y sólo si:

- p contiene una cadena $i \rightarrow m \rightarrow j$ ó $i \leftarrow m \rightarrow j$ tal que m está en Z .
- p contiene una cadena $i \rightarrow m \leftarrow j$ tal que m no está en Z y tampoco ninguno de sus descendientes.

Como consecuencia de la aplicación del criterio de *d-separación*, se establecen dos condiciones que, independientemente una de la otra, son necesarias y suficientes para que G represente a P y, por tanto, G sea una red bayesiana. La primera de ellas, denominada condición de Ordenación de Markov, establece que G es compatible con P cuando, condicionada a sus PA en G , cada variable es independiente de todos sus predecesores en cierto orden de las variables que concuerda con las flechas en G . La segunda, conocida como condición de Parentales de Markov, determina que G es

compatible con P si, condicionadas a sus PA , todas las variables son independientes de todos sus no-descendientes en G (Pearl, 2009).

3.3.3. Métodos para resolver probabilísticamente una red bayesiana

Desde un punto de vista conceptual, la determinación de la probabilidad conjunta P de la red bayesiana es un problema trivial, ya que bastaría con aplicar la Ec. 14. Sin embargo, desde un punto de vista práctico, el reto radica en diseñar procedimientos de cálculo computacionalmente eficientes que permitan calcular dicha probabilidad, considerando el nivel de representación o detalle definido por el grafo G .

Los distintos algoritmos propuestos para realizar cálculos probabilísticos en redes bayesianas están basados en la noción de propagación de la evidencia (Spiegelhalter *et al.*, 1993). Así pues, una vez que la red bayesiana ha sido construida y las distribuciones condicionadas han sido aprendidas a partir de los datos o definidas por los expertos, la distribución de probabilidad conjunta P puede ser calculada. Cuando se dispone de cierta evidencia sobre las variables de la red, estos nuevos valores de las variables aleatorias son utilizados para actualizar el cálculo de la probabilidad de la red.

Los algoritmos propuestos para el cálculo probabilístico en redes bayesianas pueden ser clasificados en dos grupos, los algoritmos exactos y los algoritmos aproximados (Charniak, 1991). Los primeros calculan la probabilidad P sin otro error que el resultante del redondeo producido por las limitaciones de cálculo del ordenador. Estos algoritmos se basan en las nociones de condicionamiento (Dechter, 1999; Lauritzen y Spiegelhalter, 1988; Zhang y Poole, 1994) y de eliminación (Darwiche, 2001; Jensen y Nielsen, 2007; Pearl, 1988) para simplificar las conexiones de los nodos de la red y reducir la complejidad computacional del proceso (Darwiche, 2010). Los segundos utilizan distintas técnicas para obtener valores aproximados de P , ya que en redes de un cierto tamaño esta tarea se convierte en un problema NP-

complejo (Cooper, 1990). Entre estos se encuentran aquellos algoritmos que permiten realizar cálculos en redes con ciclos (Weiss, 2000) o en redes con variables latentes (Jaakkola, 2000), y aquellos que permiten evaluar redes con variables continuas mediante técnicas de muestreo estocástico como la simulación Monte Carlo (Cano *et al.*, 1996; Fung y Chang, 1990; Gilks *et al.*, 1995), entre los que cabría incluir el procedimiento de cálculo desarrollado en este investigación.

Los algoritmos aproximados se sirven de la simulación para generar muestras aleatorias y calcular las probabilidades marginales de los nodos o variables deseadas de la red bayesiana, utilizando para ello distintas funciones de pesos. El motor de Simulación Monte Carlo (MSMC), que se utiliza en el procedimiento de cálculo propuesto en esta investigación, se comporta de forma análoga a estos algoritmos aproximados y permite explorar todo el universo de valores de la variable aleatoria correspondiente y, por tanto, evitar la introducción de otra forma de imprecisión o vaguedad en el modelo (Chen y Pollino, 2012). Sin embargo, la evaluación de las relaciones causales mediante el uso del MSMC sólo es posible cuando dichas relaciones son de tipo probabilístico o de tipo algebraico. En el caso que las relaciones sean de tipo difuso, resulta necesario emplear de un Motor de Inferencia Difusa (MID) capaz de evaluar este tipo de relaciones.

3.3.4. Construcción de redes bayesianas causales y evaluación de intervenciones

La construcción del DAG de una red bayesiana mediante el uso de información causal presenta varias ventajas respecto a aquellos modelos construidos en base información puramente asociacional (Pearl, 2009). Por una parte, los fundamentos causales para la formulación del modelo son más inteligibles para la mente humana, de tal forma que, la determinación de las relaciones de independencia condicional, mediante el criterio de *d-separación*, resultan más accesible y también más fiables, mientras que las relaciones de dependencia que no son soportadas mediante vínculos causales son consideradas espurias. Por otra parte, las redes bayesianas causales

constituyen modelos flexibles sobre el fenómeno de estudio, ya que son capaces de representar cambios espontáneos o externos en dicho fenómeno a través de la reconfiguración de su topología, es decir, mediante la adición o eliminación de variables y/o conexiones entre los nodos.

Esta flexibilidad de las redes bayesianas se basa en dos conceptos interrelacionados como son la modularidad y la intervención. Las condiciones de Markov determinan que cada relación entre un nodo y sus padres markovianos, $PA_i \rightarrow X_i$, representa un mecanismo autónomo y estable que puede ser manipulado sin afectar al resto de relaciones de la red (modularidad). Al mismo tiempo, esta modularidad permite determinar cómo valores concretos (intervenciones) de estos nodos parentales afectan a la probabilidad marginal sus nodos hijos dada la relación de causalidad existente entre ellos (Pearl, 2009).

En el procedimiento de cálculo que se desarrolla en el epígrafe siguiente, cada una de las m simulaciones realizadas es considerada como una intervención. Así pues, el MSMC genera valores concretos para las variables raíz de la red bayesiana y, mediante la evaluación de las relaciones causales definidas a lo largo de toda la red, estima el valor de los correspondientes nodos descendientes. De esta forma, se consigue un modelo mucho más informativo para la toma de decisiones que, además de determinar la probabilidad de los eventos y sus modificaciones como consecuencia de observaciones posteriores, permite conocer como los cambios producidos en dichas probabilidades cambiarían los resultados. Este último tipo de análisis puede resultar de especial interés en el diseño de políticas o en distintos ámbitos de gestión, puesto que reduce la incertidumbre de los agentes decisores mediante la generación de información útil para la resolución del problema objeto de estudio.

3.4. La propuesta metodológica como procedimiento de cálculo de la teoría T

En este apartado se desarrolla el procedimiento de cálculo que permite estimar la distribución de probabilidad conjunta P de la red bayesiana diseñada. Este procedimiento se desarrolla mediante la hibridación de un motor de Simulación Monte Carlo (MSCM) y un Motor de Inferencia Difusa (MID), permitiendo así la evaluación de las relaciones inciertas que caracterizan el fenómeno objeto de estudio. El MSCM genera muestras aleatorias de datos a partir de las distribuciones especificadas para las variables del modelo, cuando estas son conocidas, y evalúa las relaciones causales, igualmente conocidas o con forma funcional definida. El MID se incluye en este sistema para evaluar las relaciones de tipo difuso, obteniendo así un resultado que puede ser nuevamente procesado por el MSCM. Para disminuir la subjetividad en la definición de los parámetros de este procedimiento se incorpora un Algoritmo Evolutivo Multi-objetivo (AEMO), completando así el procedimiento de cálculo propuesto para la teoría T sobre la incertidumbre que se pretende desarrollar en esta investigación.

3.4.1. El motor de simulación Monte Carlo para la evaluación de relaciones causales

En el ámbito de la modelización de sistemas, la simulación es un enfoque que permite el estudio de la realidad de una forma eminentemente experimental, mediante la creación de modelos que reproducen el comportamiento del fenómeno bajo estudio (White y Ingalls, 2009). Entre las distintas técnicas de simulación⁴, la simulación Monte Carlo, resulta la más adecuada para desarrollar el procedimiento de cálculo deseado, puesto que su estrategia de implementación permite utilizar toda la información contenida en cada una de las variables estocásticas del modelo bayesiano.

⁴ Simulación de sistemas continuos, simulación de eventos discretos, simulación basada en agentes, simulación Monte-Carlo y simulación híbrida (White e Ingalls, 2009).

En este sentido, la simulación Monte Carlo está basada en la obtención de muestras artificiales de datos, y en el análisis estadístico para la obtención de resultados (Raychaudhuri, 2008). El funcionamiento de un MSMC puede explicarse de manera análoga al de un modelo matemático que expresa el comportamiento de un determinado sistema. De forma que, asignando valores a los inputs de este modelo, estos serán procesados mediante las formulas matemáticas, si existen, obteniendo así valores para los outputs de dicho modelo.

En el caso del MSMC incluido en el procedimiento de cálculo, los inputs son las variables aleatorias de la red, definidas mediante las correspondientes distribuciones de probabilidad, a partir de las cuales se generan muestras artificiales de datos. Estas muestras son procesadas según las relaciones definidas, algunas de ellas de naturaleza difusa -como se verá más adelante, obteniendo en cada simulación un valor aleatorio para cada uno de los outputs del modelo. Mediante la iteración de este proceso se obtiene un conjunto de valores para cada uno de los outputs del modelo. Dado que las variables inputs son aleatorias, el resultado de la red bayesiana, así como los resultados intermedios de las distintas relaciones que definen el modelo, son también variables aleatorias (Nakayama, 2008).

Posteriormente, la aplicación de técnicas de análisis estadístico permite obtener una aproximación a la probabilidad conjunta P de la red diseñada, y caracterizar la variabilidad de los valores obtenidos para los outputs, logrando así una medida de rendimiento del modelo (Raychaudhuri, 2008). Para ello, generalmente, se calcula un estimador del parámetro poblacional (output) que se desea computar, normalmente la media muestral de tamaño n (Ecs. 15 y 16), y un intervalo de confianza para el valor de alfa elegido (Ec. 17).

$$\mu = E[X] \quad (15)$$

$$\bar{X}(n) = \frac{1}{n} \sum_{i=1}^n X_i \quad (16)$$

$$E = \frac{t_{n-1, 1-\alpha/2} \sigma}{\sqrt{n}} \quad (17)$$

Sin embargo, existen determinadas matizaciones en estos cálculos según el tipo de proceso estocástico modelado, el cual es determinado por la naturaleza de las variables aleatorias (Law, 2007). Si las variables aleatorias son discretas, el proceso de simulación se califica como transitorio (*transient simulation*) y la medida de rendimiento elegida puede ser calculada mediante las ecuaciones anteriores (Ecs. 15 a 17). En el caso de variables aleatorias continuas, como se plantea en esta investigación, el proceso de simulación es calificado como estacionario (*steady-state simulation*) y resulta necesario emplear una estrategia algo diferente para el cálculo del intervalo de confianza (Nakayama, 2008).

Los datos obtenidos en un proceso de simulación estacionario no pueden ser considerados como independientes e idénticamente distribuidos (i.i.d.), lo que impide utilizar la varianza muestral (σ^2) como estimador de la varianza poblacional. Para salvar este obstáculo existen distintas técnicas como el bootstrapping, el algoritmo de réplica única, la media por lotes, los métodos espectrales, entre otros.

De entre las técnicas para la estimación de la varianza, el “bootstrapping” es una de las más utilizadas debido a que la actual capacidad de computación permite su implementación de forma sencilla (Law, 2007). Bajo esta técnica, el número de simulaciones m que se realiza en un único experimento, se divide en b experimentos con un menor número de simulaciones. Cada uno de estos experimentos recibe el nombre de lote. El número de lotes de un experimento debe oscilar entre $10 \leq b \leq 30$, aunque puede ser superior (Nakayama, 2008). Si cada experimento se realiza sin solapamiento en la generación de números pseudoaleatorios y con las mismas

condiciones iniciales, los lotes serán i.i.d., solventando de esta forma el escollo existente para el cálculo de la varianza. No obstante, al reducir el número de simulaciones por lote, los valores obtenidos estarán influenciados por las condiciones iniciales del sistema, provocando que las medias de cada lote estén sesgadas. Esto exige eliminar un determinado número de simulaciones iniciales en cada lote (Nakayama, 2008). Una vez hecho esto, resulta posible calcular la media para cada uno de los lotes \bar{y}_b (Ec. 18), la media global \bar{y} (Ec. 19) y la desviación estándar σ (Ec. 20), y, con estos valores, el error estadístico E (Ec. 21)⁵, obteniendo así una medida de rendimiento aplicable en el caso de procesos de simulación de tipo estacionario:

$$\bar{y}_b = \frac{1}{n_b} \sum_{k=n_b(b-1)+1}^{bn_b} y_k \quad (18)$$

$$\bar{y} = \frac{1}{b} \sum_{l=1}^b \bar{y}_l \quad (19)$$

$$\sigma = \sqrt{\frac{1}{b-1} \sum_{l=1}^b (y_l - \bar{y}_l)^2} \quad (20)$$

$$E = \frac{t_{b-1, 1-\alpha/2} \sigma}{\sqrt{b}} \quad (21)$$

siendo $k=1,2,\dots,m$ el número de simulaciones, $l=1,2,\dots,b$ el número de lotes y $t_{b-1, 1-\alpha/2}$ una t-Student con $b-1$ grados de libertad.

⁵ Esta formulación hace referencia a los procesos en los que el tiempo es un valor discreto, su desarrollo para valores continuos puede verse en (Nakayama, 2008).

3.4.2. El motor de inferencia difusa para la evaluación de relaciones causales difusas

3.4.2.1. Definición y funcionamiento de un motor de inferencia difusa convencional

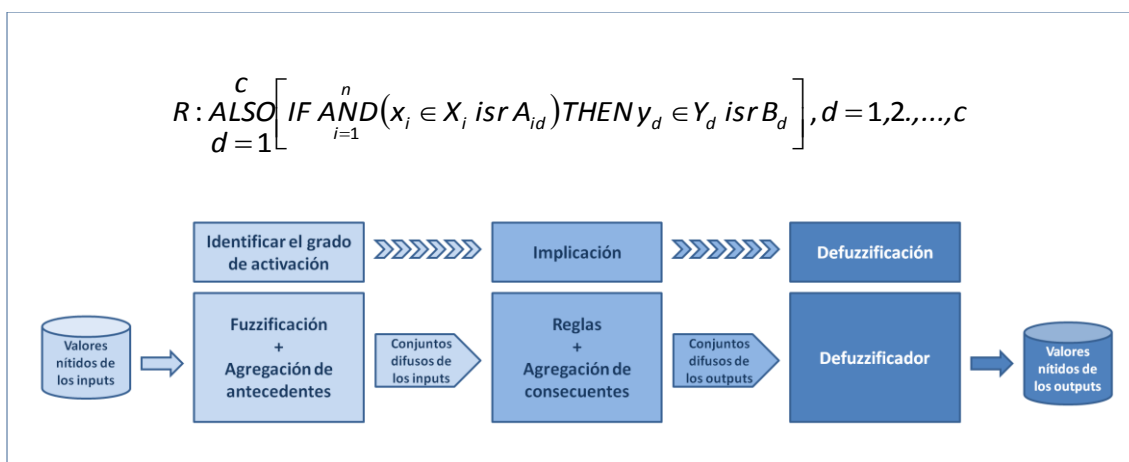
Un motor de inferencia difusa (MID) es un esquema de computación basado en la lógica difusa que permite realizar inferencia, es decir, deducir matemáticamente el output, normalmente único, de una relación difusa a partir del conjunto de reglas difusas que la definen (Celikyilmaz y Türksen, 2009). Así pues, dada una relación difusa $Y=u(X_i)$ definida mediante un conjunto de reglas R , un MID permite la obtención de un valor nítido (numérico) de y mediante la aplicación de los métodos de inferencia difusa sobre los valores observados de las variables inputs $x_i = (x_1, x_2, \dots, x_n)$.

Para poder llevar a cabo el proceso de inferencia difusa, un MID integra cuatro bloques de elementos (Jang, 1993): i) una base de conocimiento, formada a su vez por una base de reglas difusas del tipo *IF-THEN* y una base de datos de valores lingüísticos o conjuntos difusos empleados en las reglas difusas; ii) una interfaz de fuzzificación, que asigna a cada valor de la variable de entrada un grado de pertenencia sobre los conjuntos difusos definidos; iii) un mecanismo de inferencia, que ejecuta las operaciones de inferencia sobre las reglas difusas, relacionando los conjuntos difusos de las variables de entrada (antecedes de la regla) con los conjuntos difusos de la variable de salida (consecuente de la regla); y iv) una interfaz de defuzzificación, que permite obtener un valor nítido del output a partir del conjunto difuso asignado a dicho output mediante el mecanismo de inferencia.

Según la estructura de la base de reglas difusas, se pueden distinguir dos grandes familias de MID, el modelo *Sugeno* (Takagi y Sugeno, 1985) y el modelo *Mamdani* (Mamdani y Assilian, 1975). En el primero de ellos, el antecedente de la regla es difuso pero el consecuente presenta una naturaleza funcional, por lo que no es necesario que el MID incluya una interfaz de defuzzificación. En el segundo, los antecedentes y consecuentes de las reglas emplean valores difusos. La elección entre

ambos modelos radica en establecer un compromiso entre legibilidad y precisión, mientras que los modelos tipo *Sugeno* suelen producir soluciones más precisas, los modelos *Mamdani* permiten una mayor interpretabilidad de las reglas difusas. En la Figura 3 se muestra, de forma esquemática, el proceso de inferencia en un MID tipo *Mamdani*, por ser este el modelo elegido en la presente investigación. Este proceso consta de cinco fases:

Figura 3. Estructura de un MID tipo Mamdani



Fuente: Adaptado de (Celikyilmaz y Türksen, 2009).

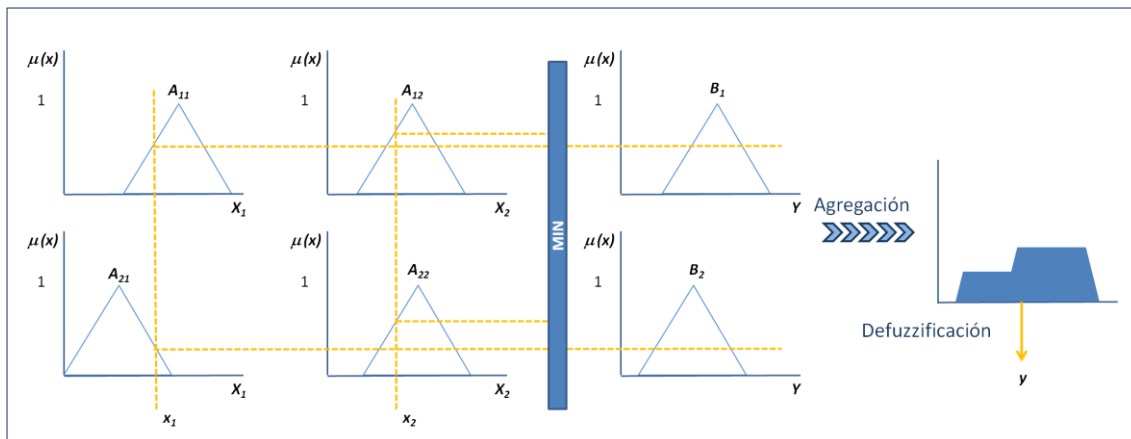
- Fuzzificación: Esta fase consiste en calcular el grado de similaridad entre los valores de las variables de entrada $X_i = (X_1, X_2, \dots, X_n)$ y las etiquetas lingüísticas A_{id} de los antecedentes de las c reglas que definen la relación difusa. Cuando los valores de las entradas son valores concretos, estos grados de similaridad se corresponden con los grados de pertenencia asociados a los valores de entrada. Para calcular estos valores basta con evaluar las funciones de pertenencia de las etiquetas lingüísticas $\mu_{A_{id}}$, es decir, $\mu_{A_{id}}(x_i): X_i \rightarrow [0,1], \forall i = 1, 2, \dots, n, \forall d = 1, 2, \dots, c$.
- Agregación de antecedentes: las funciones de pertenencia asignadas en el paso anterior son agregadas, obteniendo así el grado de disparo o grado de activación de la regla. La operación de agregación se realiza mediante la aplicación de operadores difusos, siendo el más común la t -norma usando en conector *MIN*, denotado por su análogo natural *AND*. Por tanto el grado de disparo del

antecedente de la regla d puede ser calculado como

$$\tau_d(x_i) = \underset{i=1}{\text{AND}}^n \mu_{A_{id}}(x_i), \forall i = 1, 2, \dots, n, \forall d = 1, 2, \dots, c.$$

- Implicación: A partir del grado de activación identificado en la fase anterior, mediante el uso del operado implicación, se obtiene la conclusión parcial de la regla d . Para ello, se identifica el conjunto difuso B_d del consecuente de la regla y su correspondiente función de pertenencia y grado de pertenencia, de tal forma que, $\mu_{B_d}(y_d): Y_d \rightarrow [0,1], \forall i = 1, 2, \dots, n, \forall d = 1, 2, \dots, c$.
- Agregación de los consecuentes: dado que en el paso anterior se obtiene una función de pertenencia para el consecuente de cada regla que define la relación difusa, las distintas funciones de pertenencia son agregadas mediante el operador *ALSO*, es decir, $\mu'_B(y) = \underset{d=1}{\text{ALSO}}^c \mu_{B_d}(y_d), \forall d = 1, 2, \dots, c$ obteniendo así la conclusión global para la relación definida.
- Defuzzificación: en la última fase del proceso de inferencia se obtiene un valor nítido y para cada relación difusa $Y=u(X_i)$. Para ello se pueden aplicar distintos métodos de defuzzificación (Celikyilmaz y Türksen, 2009), siendo uno de los más usado el centro de área o centro de gravedad (Cox, 2005). Este método utiliza el centro de área resultante de la asociación de los consecuentes \hat{y} como valor representativo de y (Ec. 22). Su representación gráfica aproximada puede verse en la Figura 4:

$$y \approx \hat{y} = \frac{\int_y y \cdot \mu'_B(y)}{\int_y \mu'_B(y)} \quad (22)$$

Figura 4. Sistema de Inferencia Difusa de Mamdani

Fuente: Elaboración propia.

3.4.2.2. Limitaciones en el diseño de un MID

El diseño de un MID presenta dos retos asociados a la definición de la base de conocimiento a priori y a la carga computacional derivada de la evaluación de las reglas. La definición de la base de conocimiento a priori exige de la formulación de todas las reglas difusas que describen el comportamiento del sistema. En sistemas compuestos por gran número de variables, la cantidad de reglas necesarias para su comportamiento presenta un crecimiento exponencial que dificulta su definición en base al conocimiento (Giiven y Passino, 2001). Así pues, el número total de reglas r_{tot} que define el sistema objeto de estudio es una función exponencial del número de conjuntos difusos e seleccionados para los n inputs del modelo y del número de relaciones difusas h (Gegov y Gopalakrishnan, 2008):

$$r_{tot} = he^n \quad (23)$$

Como consecuencia de lo anterior, el MID debe realizar un elevado número de operaciones, incrementando la complejidad computacional del proceso de inferencia. Este hecho, conocido como “maldición de la dimensionalidad” (Giiven y Passino, 2001),

ha dado lugar a distintas propuestas para reducir el número de reglas a procesar por el MID. Para ello se han implementado algoritmos genéticos (Ishibuchi *et al.*, 1995), redes neuronales (Halgamuge y Glesner, 1994) o distintas variedad de técnicas de clustering (Chiu, 1997) que intentan seleccionar aquellas reglas que tienen mayor relevancia en el proceso de inferencia de resultados. Otras propuestas intentan reducir el número de reglas mediante la modificación de los conectivos lógicos en la estructura de las reglas *IF-THEN* (Combs y Andrews, 1998), o mediante la sustitución del sistema difuso original, caracterizado por el elevado número de inputs, por un conjunto de sistemas difusos de menor dimensión enlazados de forma jerárquica (Raju *et al.*, 1991).

3.4.2.3. Diseño de un motor de inferencia difusa para la evaluación de relaciones causales en la red bayesiana (MID-RB)

3.4.2.3.1. Automatización del proceso de creación y evaluación de reglas: Parámetros del MID-RB

El MID-RB empleado en esta investigación (García-Alonso *et al.*, 2012), afronta las limitaciones señaladas anteriormente mediante el desarrollo de un procedimiento en el que sólo se diseñan y evalúan aquellas reglas asociadas al valor que presentan los inputs del sistema. Este proceso se lleva a cabo de forma automática mediante el establecimiento de un conjunto de parámetros definidos en base al conocimiento. Así pues, una vez identificadas las relaciones difusas y el universo de discurso de las variables implicadas en dichas relaciones, resulta necesario definir tres grupos de parámetros (García-Alonso y Pérez-Alcalá, 2008):

- *Peso* o nivel de relevancia: el peso indica la importancia relativa de cada variable input sobre el output de una relación difusa. El peso de cada variable, denotado como w_{id} , es asignado mediante distribución de probabilidad que refleja el conocimiento experto sobre dicho peso.

- *Tipo* de la relación difusa: el tipo es una categoría definida por los diseñadores del MID que hace referencia a la relación entre el antecedente y el consecuente de la regla difusa. El tipo permite reflejar la forma imprecisa en la que el experto puede definir la influencia que ejercen los inputs sobre el output de una relación difusa. Estas categorías deben ser definidas de forma que resulten entendibles para el experto y así este pueda seleccionar aquella que mejor se ajuste a la definición de la relación difusa. En la propuesta metodológica se especifican cinco tipos:
 - Neutra: la combinación de inputs ejerce una influencia neutra, positiva o negativa sobre el output. Se asociaría a declaraciones causales del tipo *“las remesas registradas **dependen positivamente** del stock de migrantes y negativamente de los costes de transferencia y de los tipos de cambio”* (Freund y Spatafora, 2008).
 - Positivo/Negativo: la combinación de inputs ejerce una influencia positiva/negativa más que proporcional sobre el output. Se asociaría a declaraciones causales del tipo *“las remesas disminuyen a un ritmo **más rápido** cuando los migrantes tienen lazos débiles con sus países de origen”* (Amuedo-Dorantes y Pozo, 2006b).
 - Muy positivo/Muy negativo: la combinación de inputs ejerce una influencia positiva/negativa mucho más que proporcional sobre el output. Se asociaría a declaraciones causales del tipo *“Los hogares pobres en comunidades con grandes redes sociales de migrantes son **mucho más proclives** a migrar, y por tanto a enviar remesas, que los hogares pobres con redes migratorias pequeñas”* (McKenzie, 2005).

- *Intensidad* de la relación difusa: la intensidad es otra categoría definida por los diseñadores que hace referencia a la relación entre el antecedente de una regla difusa y el grado de cumplimiento o de verdad de dicha regla. Dada una relación difusa, los expertos pueden identificar zonas en el universo de discurso de las variables input, en las cuales el comportamiento definido por la regla puede ser

discutible o ambiguo, lo que afecta al grado de verdad o nivel de cumplimiento de la regla. Así pues, la intensidad de la relación difusa permite incorporar al sistema difuso aquellos matices, señalados por los expertos, sobre el grado de cumplimiento de las correspondientes reglas difusas. Al igual que en el caso anterior, se han definido cinco categorías para este parámetro que, asimismo, deben ser entendibles por el experto:

- Neutra: no se establece ningún matiz sobre el grado de cumplimiento de las reglas difusas.
- Creciente / Decreciente: la regla define mejor el comportamiento de la relación difusa a medida que los inputs de dicha relación presentan valores más elevados o más bajos en los correspondientes universos de discurso y, por tanto, el grado de cumplimiento de las reglas difusas será mayor en estas zonas.
- Centrado creciente / Centrado decreciente: la regla define mejor el comportamiento de la relación difusa a medida que los inputs de la misma toman valores centrales en los correspondientes universos de discurso y, por tanto, el grado de cumplimiento de las reglas difusas será mayor en esta zona central.

Por último, cabe señalar que, cuando el grado de cumplimiento de una regla es menor que una determinada probabilidad fijada en base al conocimiento experto, la selección del conjunto difuso para el output de dicha regla se realiza mediante el MSMC.

3.4.2.3.2. Funcionamiento del MID-RB

Como ya se ha señalado el MSMC incorpora un MID, basado en una estructura tipo *Mamdani*. El MID-RB diseña, instancia y evalúa de forma automática solo aquellas reglas difusas que se instancian con los valores de los inputs generados por el MSMC o que son resultado de la evaluación de otra relación causal, tanto de tipo algebraico

como de tipo difuso.

El MID-RB calcula la función de pertenencia B_d para cada output y_d mediante una media ponderada (Ec. 24)(García-Alonso y Pérez-Alcalá, 2008):

$$B_d = \text{round} \left[\left(\frac{\sum_{i=1}^n w_{id} A_{id}}{\sum_{i=1}^n w_{id}} \right) + Q + D \right] \quad (24)$$

siendo $d= 1,2,\dots,c$ el número de reglas que define la relación causal difusa; $i= 1,2,\dots,n$ el número de inputs en dicha relación; A_{id} la función de pertenencia y w_{id} el peso o importancia relativa de cada variable input x_i en el output y_d ; B_d la función de pertenencia para el output Y , ya que cada regla d tiene un único output y_d ; Q y D son los dos parámetros que reflejan respectivamente el *tipo* y la *intensidad* de la correspondiente relación difusa.

El parámetro Q permite que, durante el proceso de inferencia, se consideren la imprecisión o vaguedad con la que los expertos han podido definir la relación difusa correspondiente (*tipo*). Por su parte, el parámetro D permite computar los matices establecidos sobre el grado de cumplimiento de las reglas, denotado como fd , que definen las relaciones difusas (*intensidad*). Q y D son calculados a partir del *tipo* y la *intensidad* de la relación difusa respectivamente, según las categorías establecidas por los expertos durante la fase de definición de los parámetros de la red bayesiana.

Así pues, Q es un número entero seleccionado a partir de i) las categorías definidas para el *tipo* de relación difusa y ii) del número de funciones de pertenencia e definidas para las variables input de dicha relación, tal que $|Q| \leq (e-1)/2$. El número y la forma de las funciones de pertenencia e son definidas por los diseñadores del sistema. Si consideramos las cinco categorías definidas con anterioridad para el *tipo* de relación difusa (neutra, positivo, muy positivo, negativo, muy negativo), Q tomará valores

entre $[-2,2]$. Esto implica un desplazamiento de la función de pertenencia del output hacia a la izquierda o hacia la derecha. Cuando $Q=0$ (*tipo=neutra*) la función de pertenencia del output, determinada mediante (Ec.24), queda inalterada.

El segundo parámetro, D , también es un número entero tal que $|Q|+|D|\leq(e-1)/2$, seleccionado de manera aleatoria dentro de un intervalo de número enteros $[-v,v]$, incluido el cero. El valor de v depende del valor de fd , de tal forma que cuanto menor sea fd , mayor es el valor de v y, por tanto, también el rango del intervalo $[-v,v]$. Además el valor de v debe cumplir que $|v|\leq(e-1)/2$. El parámetro D también comporta un de desplazamiento de la función de pertenencia del output, calculada conforme a (Ec.24).

Como resultado de introducir ambas distorsiones, Q y D , en el cálculo de la función de pertenencia B_d , ésta estará sesgada, es decir, desplazada a la izquierda o derecha respecto a la posición que hubiera tenido sobre el espacio de discurso de la variable output de la relación difusa sin considerar dichos parámetros (situación completamente neutra). De esta forma se consigue reflejar de forma más precisa el estado actual de conocimiento sobre el fenómeno modelado.

La determinación de la función de pertenencia del output B_d mediante el MID-RB se inicia con la fuzificación de los valores de los inputs x_i . Estos valores son generados por el MSMC y/u obtenidos de la evaluación de otra relación causal, permitiendo identificar las funciones de pertenencia A_{id} y calcular los grados de pertenencia correspondientes $\mu_{A_{id}}(x_i)$ ($i=1,2,\dots, n$ y $d=1,2, \dots,c$). A continuación, el MSCM estima los pesos de los inputs w_{id} ($i=1,2,\dots, n$ y $d=1,2,\dots,c$) a partir de las correspondientes distribuciones estadísticas. Una vez hecho esto, el MID-RB calcula el grado de cumplimiento fd de las reglas instanciadas por cada combinación de funciones de pertenencia A_{id} , según la *intensidad* definida por el experto. El fd obtenido se compara con un número pseudo-aleatorio generado por la computadora, de tal forma que cuando fd es menor que dicho numero, el parámetro D es

determinado de forma aleatoria en un rango de número enteros $[-v, v]$. Posteriormente, para la misma combinación de funciones de pertenencia A_{id} , el MID-RB determina B_d , como si la categoría seleccionada para el *tipo* de relación fuera neutra ($Q=0$). Teniendo en cuenta la categoría seleccionada por el experto para el *tipo* de relación y la B_d obtenida en el paso inmediatamente anterior, se calcula el parámetro Q . Los valores obtenidos para Q y D son empleados para calcular la función de pertenencia B_d conforme a (Ec. 24). Finalmente, la aplicación del método del centro de gravedad (*Product-sum gravity method*) (Cox, 2005) permite obtener el valor nítido del output y_d .

3.4.3. Validación de los resultados

La validación del sistema diseñado se lleva a cabo mediante la comparación de los resultados obtenidos con los datos reales disponibles, teniendo en cuenta que es posible conocer, al menos de forma fragmentaria, los valores de varias variables. Para ello, cuando es posible, se calcula el error cuadrático medio (ECM) para las $k=1,2,\dots,m$ simulaciones:

$$ECM = \left[\frac{1}{m-1} \sum_{k=1}^m (y_k - y)^2 \right] \quad (25)$$

3.4.4. El Algoritmo Evolutivo Multi-objetivo para la determinación de los pesos de las variables input en las relaciones difusas

3.4.4.1. Elementos de un Algoritmo Evolutivo Multi-objetivo

Desde el trabajo inicial de Schafer (1985), los Algoritmos Evolutivos Multi-Objetivo (AEMO) se han consolidado como una potente y robusta metaheurística de

optimización, dando lugar a una amplia variedad de enfoques (Zhou *et al.*, 2011). Un AEMO es una clase de algoritmo genético, inspirado por tanto en los mecanismos de selección de la Naturaleza, aplicado a la resolución de problemas multi-objetivo.

Los problemas multi-objetivo, también denominados, problemas de optimización con criterios múltiples o de optimización vectorial, pueden ser definidos como *“el problema de encontrar un vector de variables de decisión que satisfaga las restricciones y optimice una función vectorial cuyos elementos representen las funciones objetivo. Estas funciones forman una descripción matemática de los criterios de desempeño que generalmente están en conflicto entre sí. Por lo tanto, el término “optimizar” significa encontrar una solución tal que proporcione valores para todos los objetivos que resulten aceptables para el diseñador”* (Osyczka, 2002).

La definición anterior permite establecer como principales elementos de un AEMO los siguientes:

- Un vector de variables de decisión $\mathbf{x}=[x_1, x_2, \dots, x_n] \in \Omega$, siendo n el número de variables y Ω el conjunto de todas las soluciones factibles que satisfacen un conjunto de restricciones.
- Un conjunto de restricciones, que define la región de soluciones factibles del problema, incluyendo m restricciones de desigualdad, $g_i(\mathbf{x}) \leq 0, \forall i = 1, 2, \dots, m$, y p restricciones de igualdad, $h_j(\mathbf{x}) \leq 0, \forall j = 1, 2, \dots, p$, de tal forma que, normalmente, pero no siempre, $p < n$ para evitar sobre-restricciones.
- Un vector de funciones objetivo definidas sobre las variables de decisión $\mathbf{f}(\mathbf{x})=[f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_k(\mathbf{x})]$. La relación entre estas funciones suele ser de conflicto, es decir, no existe ningún vector factible de variables de decisión $\mathbf{x} \in \Omega$ que provoque un incremento en alguna de las funciones objetivo, sin causar simultáneamente un decremento por lo menos en otra (eficiencia de Pareto).

- Un algoritmo genético, como metodología de optimización, que pretende encontrar una aproximación al conjunto de soluciones eficientes, dentro de Ω , para todas las funciones objetivo del problema.

3.4.4.2. Funcionamiento de un AEMO

En un AEMO la búsqueda de soluciones está guiada por el concepto de Optimalidad de Pareto (Fonseca y Fleming, 1995). Dado el vector de funciones objetivo $\mathbf{f}(\mathbf{x})=[f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_k(\mathbf{x})]$, el vector de soluciones $\mathbf{x} \in \Omega$ es un óptimo de Pareto, también llamado solución no dominada, si no existe otro vector $\mathbf{x}' \in \Omega$ tal que, $f_i(\mathbf{x}') \leq f_i(\mathbf{x})$, $\forall i=1,2,\dots,k$, y $f_i(\mathbf{x}') < f_i(\mathbf{x})$ para al menos un valor de i . La no dominancia puede ser establecida tanto en sentido débil como en sentido estricto. En el primer caso, se dice que un vector $\mathbf{x} \in \Omega$ es una solución débilmente no dominada si no existe otro vector $\mathbf{x}' \in \Omega$ tal que, $f_i(\mathbf{x}') < f_i(\mathbf{x})$, $\forall i=1,2,\dots,k$. En el segundo caso, el vector $\mathbf{x} \in \Omega$ es una solución estrictamente no dominada si no existe otro vector $\mathbf{x}' \in \Omega$ tal que, $f_i(\mathbf{x}') \leq f_i(\mathbf{x})$, $\forall i=1,2,\dots,k$, y $f_i(\mathbf{x}') < f_i(\mathbf{x})$ para al menos un valor de i .

En los problemas multi-objetivo en los que existe un cierto grado de conflicto entre los objetivos y variables continuas, suele haber un conjunto, normalmente infinito, de soluciones eficientes. Estas soluciones no dominadas son llamadas conjunto de óptimos de Pareto (P^*), de forma que $P^*:\{\mathbf{x} \in H | \neg \exists \mathbf{x}' \in H f(\mathbf{x}') \prec f(\mathbf{x})\}$. La representación gráfica de las funciones objetivo correspondientes al conjunto de óptimos de Pareto se denomina Frente de Pareto (FP^*). La obtención de la expresión analítica de FP^* es una tarea compleja y en algunos casos imposible. Sin embargo, esta puede ser aproximada mediante la generación de un número suficiente de vectores de solución y la comprobación sistemática de la no dominancia entre ellos.

En los AEMO, la determinación de la no dominancia de las soluciones se lleva a cabo mediante una función de aptitud (*fitness function*) que, en el caso de problemas multi-objetivo, debe considerar las diferentes representaciones de cada función

objetivo. En problemas con un elevado número de objetivos, el diseño de la función de aptitud puede resultar complejo, ya que su evaluación requiere de un gran esfuerzo computacional. Para ello se han desarrollado distintos enfoques como la creación de "rankings" de objetivos, la ponderación de objetivos, los modelos especiales tipo SPEA2 o la inferencia difusa (Coello *et al.*, 2007).

La estrategia de implementación de los AEMO permite encontrar múltiples soluciones no dominadas o Pareto óptimas en problemas complejos en una sola pasada, al tiempo que evitan quedar atrapado en óptimos locales (Zhou *et al.*, 2011). Esta estrategia se basa en la generación de conjuntos o poblaciones de soluciones, así como en la mejora sucesiva de dichas soluciones hasta alcanzar un determinado criterio de parada.

En los AEMO, el proceso de optimización (Figura 5) se inicia con la fijación de un número determinado semillas (población inicial) compuestas por grupos de soluciones. Estas soluciones son diseñadas de forma aleatoria o seleccionadas mediante métodos determinísticos y reciben el nombre de individuos. En cada generación las soluciones son mejoradas mediante la aplicación de operadores genéticos clásicos como la selección, la recombinación y la mutación. Además suele ser frecuente, aunque opcional, la existencia de determinado elitismo, permitiendo así la supervivencia de las mejores soluciones (Coello *et al.*, 2007).

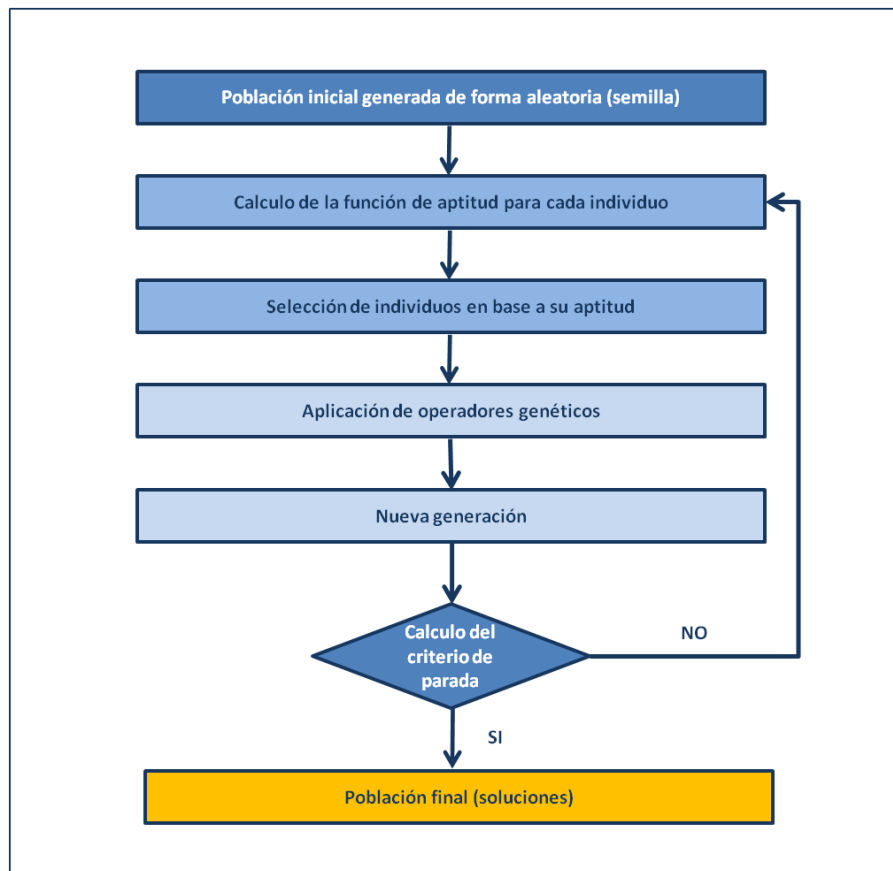
Mediante la aplicación del operador de selección, en cada generación se eliminan aquellas soluciones que presentan los peores valores en la función de aptitud definida. Las otras soluciones son preservadas para su posterior mejora mediante la aplicación del resto de operadores genéticos. La selección de soluciones se puede llevar a cabo a través de distintos métodos, entre los que destacan, la ruleta (las soluciones con mejor aptitud tienen mayor probabilidad de ser seleccionadas), el torneo (se comparan los valores de aptitud por pares de soluciones y se selecciona aquella con mejor valor) o la categoría (las soluciones se ordenan por su valor de

aptitud y se seleccionan aquellas con mejor puntuación). Las soluciones eliminadas son, normalmente, reemplazadas por una copia de las mejores soluciones preservadas de forma que se mantiene constante el número de soluciones de la población inicial en cada generación.

El operador de recombinación o cruce intercambia los componentes de dos soluciones diferentes mediante uno o dos puntos de corte, con la esperanza de crear una nueva solución que tenga una función de aptitud mejor. En cada generación, la selección de las soluciones para su recombinación es aleatoria y se ajusta mediante un porcentaje o probabilidad sobre el total de individuos de la población.

El operador mutación intenta crear mejores soluciones mediante la sustitución de uno o más componentes de una misma solución, explorando así nuevas regiones del espacio de soluciones y evitando que el algoritmo converja rápidamente a un óptimo local. La mutación puede llevarse a cabo mediante distintas técnicas entre las que destacan el cambio de uno o varios componentes de la solución en representaciones binarias o la adición de un valor aleatorio perteneciente a una distribución de probabilidad si la representación es real. Al igual que en el operador anterior, en cada generación la selección de las soluciones sobre las que se aplica la mutación se determina mediante la fijación de un porcentaje o probabilidad sobre la población de soluciones. El porcentaje de mutación suele ser menor que el cruce, ya que este operador aumenta la diversidad de la población.

Por último, la utilización en el AEMO del operador de elitismo pretende asegurar que los mejores individuos, es decir, aquellas soluciones que presentan el mejor valor de la función de aptitud, pasen a la siguiente generación sin que sean modificados por ninguno de los dos operadores anteriores. De esta forma, las mejores soluciones permanecen inalterables hasta que aparezca otra solución, a lo largo del proceso evolutivo, que las supere.

Figura 5. Funcionamiento de un AEMO

Fuente: Elaboración propia.

El proceso descrito se lleva a cabo de forma iterativa hasta que se alcanza un criterio de parada que detiene el algoritmo, obteniendo así la población final o conjunto de soluciones finales. Como criterio de parada se puede fijar un número máximo de generaciones o un número de generaciones sucesivas durante el cual no se produce una mejora significativa del valor de la función aptitud bien del mejor individuo de la población, bien de la media de aptitud de la población de soluciones. La determinación del criterio de parada, así como la de otros parámetros de un AEMO como el tamaño de la población, el número de objetivos, el tipo de selección, la existencia de elitismo, la probabilidad de los operadores genéticos, etc., son fijados inicialmente y condicionan de forma crucial el comportamiento del algoritmo, siendo este uno de los problemas más importantes que se deben afrontar.

3.4.4.3. Diseño de un AEMO para la estimación de pesos en las relaciones causales difusas de la red bayesiana (AEMO-RB)

3.4.4.3.1. Planteamiento del problema multi-objetivo en las relaciones causales difusas

La modelización de las relaciones causales difusas en una red bayesiana exige la determinación de tres grupos de parámetros, los *pesos* de las variables inputs de las relaciones difusas, el *tipo* y la *intensidad* de éstas relaciones. Como se ha mencionado anteriormente estos parámetros son determinados a partir del conocimiento existente sobre el fenómeno de estudio. Sin embargo la estimación de los *pesos* o importancia relativa de las variables inputs w_{id} puede ser también aproximada, conforme a la hipótesis establecida en esta investigación, en base a información numérica (datos) sobre la variable output de la correspondiente relación difusa.

La determinación de estos pesos w_{id} se plantea como un problema de optimización multi-objetivo, en el que las posibles combinaciones de pesos \mathbf{W}^* deben ser obtenidas mediante la optimización de dos objetivos contrapuestos, cuyos valores deben de tender a un mínimo a medida que el número de simulaciones fijado para el procedimiento de cálculo aumenta.

El primer objetivo a minimizar es el Error Cuadrático Medio (ECM), entre el valor estimado del output \hat{y}_d y su verdadero valor y_d :

$$f_{jg}^1 : \text{Min}(ECM_{jg}) : \text{Min} \left[\frac{1}{m-1} \sum_{k=1}^m (y_{jgk} - y_d)^2 \right] \quad (26)$$

El segundo objetivo, también a minimizar, es el error (E) derivado del proceso de simulación calculado según (Nakayama, 2008):

$$f_{jg}^2 : \text{Min}(E_{jg}) : \text{Min} \left[\frac{t_{b-1, 1-\alpha/2} \sigma_{jg}}{\sqrt{b}} \right] \quad (27)$$

En el AEMO-RB una solución o individuo j es un vector de pesos de las variables inputs \mathbf{w}_j que satisface la siguiente restricción (criterio de factibilidad):

$$\sum_{i=1}^n w_{ij} = 1 \quad (28)$$

siendo $i=1,2,\dots,n$ el número de inputs en una relación difusa y $j=1,2,\dots,s$ el número de soluciones o individuos en la población.

En una generación concreta g , el algoritmo analiza \mathbf{w}_{jg} , formado por un número constante de soluciones s . Este vector \mathbf{w}_{jg} es el resultado de aplicar los operadores genéticos sobre \mathbf{w}_{jg-1} hasta que se alcance el criterio de parada. Así pues, la población de s soluciones en cada generación se denota como \mathbf{W}_g .

$$\mathbf{W}_g = [(w_{11}, w_{21}, \dots, w_{n1}), \dots, (w_{1s}, w_{2s}, \dots, w_{ns})], (g=1, 2, \dots, s) \quad (29)$$

Las soluciones en cada generación g son analizadas mediante los dos objetivos anteriormente citados. Estos dos objetivos son calculados por el MSMC a partir del output y_{jgk} obtenido en cada simulación k ($k=1,2,\dots,m$) para los correspondientes valores de los inputs $i=1,2,\dots,n$. Con objeto de simplificar el proceso de validación del AEMO-RB, el número de simulaciones m se ha considerado constante, de tal forma que el proceso de simulación, que presentaba un comportamiento estacionario, pasa a tener una naturaleza transitoria. Esta simplificación supone la modificación de la formulación original del ECM (Ec. 25), de tal forma que los objetivos establecidos quedan formulados como sigue:

$$\begin{aligned}
 f_{jg}^1 &: \text{Min}(ECM_{jg}) : \text{Min} \left[\sum_{k=1}^m (y_{jgk} - y_d)^2 \right] \\
 f_{jg}^2 &: \text{Min}(E_{jg}) : \text{Min} \left[\frac{t_{b-1, 1-\alpha/2} \sigma_{jg}}{\sqrt{b}} \right]
 \end{aligned} \tag{30}$$

3.4.4.4. Diseño del AEMO-RB

La función de aptitud FF_{jg} del AEMO-RB, se ha determinado utilizando la aproximación de suma ponderada de objetivos (*Weighted Sum of Objectives*) propuesta por Das y Dennis (1997):

$$\begin{aligned}
 \text{Min } FF_{jg} &= \rho_1 \tilde{f}_{jg}^1 + \rho_2 \tilde{f}_{jg}^2 \\
 \rho_1 + \rho_2 &= 1 \\
 \rho_1, \rho_2 &\geq 0
 \end{aligned} \tag{31}$$

siendo $\tilde{f}_{jg}^1, \tilde{f}_{jg}^2$ los valores reescalados de las funciones objetivo, y ρ_1 y ρ_2 los pesos asignados a cada una de las funciones objetivo en la función de aptitud. Estos pesos son asignados de forma aleatoria por el MSMC. Así pues, el AEMO-RB debe optimizar un sistema no-lineal con un solo objetivo FF_{jg} , en el que una solución \mathbf{w}_j domina a otra solución \mathbf{w}_s si $FF_{jg}(\mathbf{w}_j) < FF_{jg}(\mathbf{w}_s)$. Para ello, se sigue el procedimiento descrito en la Figura 6.

El criterio de parada establecido para el AEMO-RB se basa en el cálculo del Error Cuadrático Medio de la función de aptitud para cada solución \mathbf{w}_j , denotado como $E_{FF_{jg}}$. Cuando el valor medio del $E_{FF_{jg}}$ del conjunto de mejores soluciones es reiteradamente menor que un valor prefijado e_{FF} , durante un número máximo de generaciones g_{max} , el algoritmo se detiene y las soluciones no dominadas u óptimas $\mathbf{W}_g^* = [\mathbf{w}_{1g}^*, \mathbf{w}_{2g}^*, \dots, \mathbf{w}_{zg}^*]$ ($e=1,2,\dots,z$) en cada generación g son guardadas en un archivo

externo denominado *SP (Solution Pool)*. En caso contrario, es decir, si no se alcanza el criterio de parada, los operadores genéticos de selección, elitismo y mutación son aplicados sobre la población en la generación g , para conseguir mejorar la función de aptitud en la próxima generación \mathbf{W}_{g+1} .

Figura 6. Procedimiento del AEMO-RB

1	Creación del fichero de soluciones <i>SP</i>	
2	For seed=1 a seed _{max} do	
3	Diseño de la población inicial \mathbf{W}_{seed}	Pesos o niveles de relevancia aleatorios
4	For $g=1$ a g_{max} do	
5	Analizar la función de aptitud de cada individuo en \mathbf{W}_g	Suma ponderada de objetivos (<i>Weighted sum of objectives</i>) Criterio de parada basados en el ECM de la aptitud de las soluciones
6	Preservar las soluciones no dominadas en <i>SP</i>	
7	Elitismo	
8	Selección por torneo binario con reemplazo	
9	Mutación	Aleatoria
10	end for	
11	end for	
W: Soluciones iniciales		SP: Solution pool
g: generaciones		ECM: Error Cuadrático Medio

Fuente: Adaptado de (García-Alonso *et al.*, in press).

En las sucesivas iteraciones del AEMO-RB y, a partir de un cierto número de generaciones, se ejecuta el operador de elitismo que preserva un número predefinido, normalmente la mejor, de aquellas soluciones con los mejores valores para la función de aptitud. Estas soluciones son almacenadas provisionalmente en el *SP* de forma que no se les aplica ningún otro operador genético.

El operador genético de selección está basado en el método estándar de torneo binario con reemplazamiento (Coello y Lamont, 2004). En este método, la selección de soluciones se efectúa mediante la comparación, por pares de soluciones, de su valor de aptitud, utilizando para ello el criterio de no dominancia tanto en sentido débil como en sentido estricto o fuerte. Las soluciones a comparar son escogidas al azar, en base a una probabilidad o porcentaje p_T prefijado, que debe ser mayor que un número pseudo-aleatorio. Como resultado de la comparación, la solución más apta es seleccionada para ser mejorada mediante la aplicación de otros operadores genéticos, mientras que la solución menos apta es eliminada y reemplazada por una nueva solución factible, consiguiendo normalmente de esta forma incrementar la aptitud global y mantener el tamaño de la población de generación en generación. Se trata por tanto de un algoritmo de censo constante. La selección por torneo tiene como ventaja su facilidad de implementación y eficiencia computacional, además no requiere reescalar los valores de aptitud y permite aumentar o reducir la oportunidad de supervivencia de las soluciones menos aptas modificando el tamaño del torneo.

Finalmente, se ha aplicado un operador de mutación que actúa sobre la población de soluciones, de forma aleatoria, cuando la probabilidad p_M es superior a otro número pseudo-aleatorio. Además, el operador mutación es aplicado de forma sistemática a todas aquellas soluciones generadas e identificadas como no factibles hasta que satisfacen la restricción relativa a los pesos (véase Ec. 30). La existencia de esta restricción hace que la aplicación del operador de recombinación o cruce carezca de sentido en el diseño del AEMO-RB. Bajo este diseño, la recombinación de soluciones daría lugar a otras soluciones o individuos que no cumplirían, en general, la restricción impuesta a los pesos. Como consecuencia de este hecho se ha optado por elevar el valor asignado a la probabilidad con la que se aplica el operador de mutación p_M .

3.5. La entropía de Shannon como medida de la incertidumbre en la teoría T

El índice de Shannon (1948), también conocido como Entropía de Shannon, se emplea para medir la incertidumbre asociada a la función de probabilidad conjunta P de la red bayesiana. La reducción de la incertidumbre mediante cualquier intervención en la red bayesiana es deseable e indica un progreso en la generación de información útil para la toma de decisiones en el ámbito correspondiente.

La funcionalidad de este índice como medida de la incertidumbre y de la información expresada mediante la Teoría de la Probabilidad ha sido constatada en numerosos estudios (Klir, 2006). En estos estudios se demuestra que dicha medida satisface los distintos conjuntos de axiomas que caracterizan a una medida probabilística de la incertidumbre.

La entropía de Shannon S sobre un conjunto X , que representa las alternativas finitas y excluyentes definidas sobre una variable aleatoria discreta objeto de estudio, y su correspondiente distribución de probabilidad $p(x)$, es definida como (Klir, 2006):

$$S(p) = -\sum_{x \in X} p(x) \log_2 p(x) \quad (32)$$

En esta medida, la probabilidad y la incertidumbre se relacionan de forma inversa, es decir, a medida que la probabilidad de una alternativa aumenta, la incertidumbre sobre la ocurrencia de dicha alternativa se reduce y, por tanto, la información proporcionada por la constatación real de dicha alternativa también es menor. A diferencia de la probabilidad, que toma valores en el intervalo $[0,1]$, la entropía toma valores en el intervalo $[0,\infty)$, siendo la unidad de medida asociada el bit.

Así pues, una variable aleatoria binaria X cuyos estados son equiprobables, tendría asociada una incertidumbre $S(X)=1$ aplicando (Ec. 32), mientras que si uno de estos estados fuera más probable, por ejemplo, $p(x=0)=0,1$ y $p(x=1)=0,9$, la incertidumbre asociada sería $S(X)=0,47$. La diferencia entre ambas situaciones apunta una reducción de la incertidumbre (o incremento de la información) en 0,53 bits.

Esta medida puede ser generalizada para n variables aleatorias *d-separadas* en una red bayesiana. Así pues, dados dos conjuntos X e Y de estados, que pueden ocurrir en sus correspondientes variables aleatorias, resulta posible definir tres formulaciones de la Entropía de Shannon (Klir, 2006) como son la Entropía simple $S(X)$, la Entropía conjunta $S(X,Y)$ y la Entropía condicionada $S(X|Y)$:

$$S(X) = - \sum_{x \in X} p_x(x) \log_2 p_x(x) \quad (33)$$

$$S(X,Y) = - \sum_{\langle x,y \rangle \in X \times Y} p(x,y) \log_2 p(x,y) \quad (34)$$

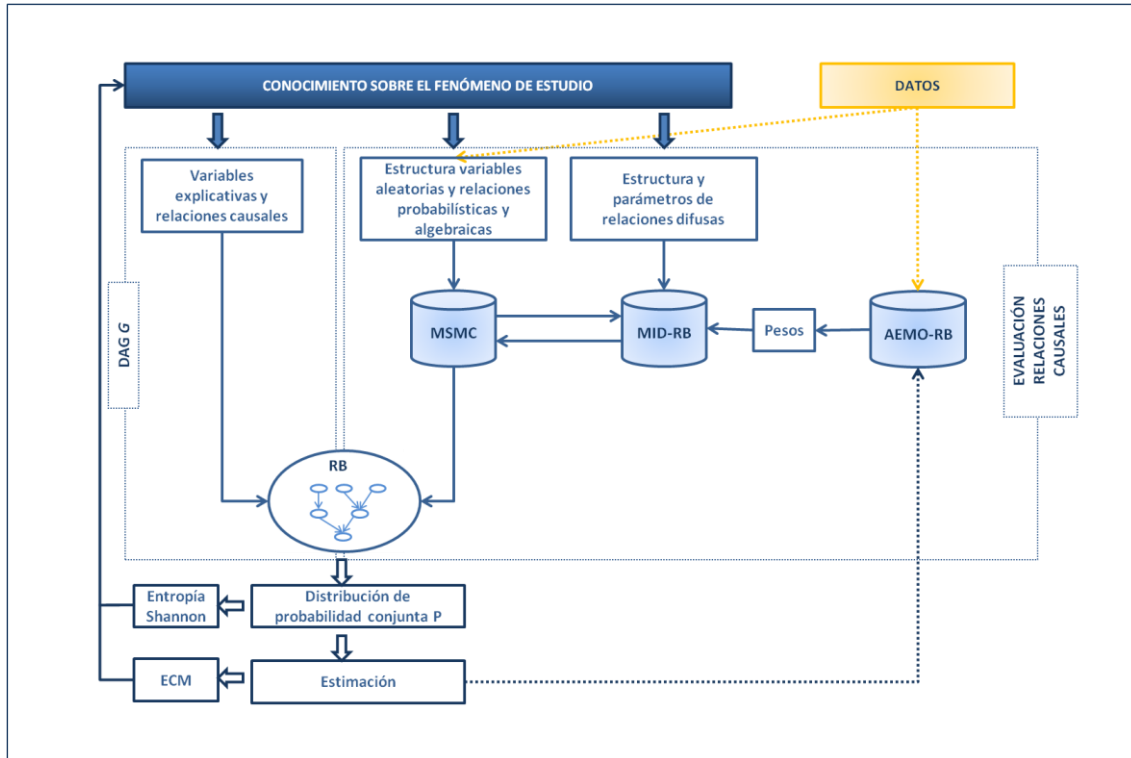
$$S(X|Y) = - \sum_{y \in Y} p_y(y) \sum_{x \in X} p(x|y) \log_2 p(x|y) \quad (35)$$

3.6. La teoría T sobre la incertidumbre: una visión de conjunto

La propuesta metodológica desarrollada a lo largo de este capítulo dará lugar al diseño de una red bayesiana causal, cuyas variables y relaciones causales serán evaluadas mediante el procedimiento de cálculo descrito con anterioridad, obteniendo como resultado la función de distribución de probabilidad conjunta P asociada a dicha red. Esta propuesta metodológica, que se muestra de forma esquemática en la Figura 7, se inicia con la identificación, a partir del conocimiento existente, de las variables

explicativas del fenómeno estudiado y las relaciones causales entre las mismas, lo que permite construir el DAG G de la red bayesiana (RB).

Figura 7. Esquema resumen de la propuesta metodológica



Fuente: Elaboración propia.

Este conocimiento permite identificar el tipo de relación causal existente, probabilística, algebraica o difusa, y, en algunos casos, las distribuciones estadísticas que describen el comportamiento de los nodos raíz de la red bayesiana. La existencia de datos sobre algunas de las variables del modelo puede emplearse también para la determinación de las distribuciones estadísticas correspondientes. De acuerdo con este conocimiento, también es posible determinar todos aquellos parámetros necesarios para la evaluación de las relaciones difusas.

La evaluación de la red bayesiana causal se lleva a cabo mediante el procedimiento de cálculo en el que el MSMC genera, cuando es necesario, muestras aleatorias para las variables raíz de la red bayesiana según las distribuciones

estadísticas correspondientes. Estos valores son procesados según las relaciones causales definidas entre estas variables raíz y los nodos hijos. Las relaciones probabilísticas son evaluadas de forma convencional, mientras que las relaciones de tipo algebraico, $Y=f(X_i)$, son procesadas por el MSMC según la forma funcional definida. En caso de ser relaciones de tipo difuso, $Y=u(X_i)$, los valores hacen que el MID-RB diseñe y evalúe las correspondientes reglas difusas, utilizando los pesos determinados por el AEMO-RB, en el caso de existir datos sobre el output de la relación difusa correspondiente, o en su defecto los generados por el MSMC. En cualquier caso, como resultado de la evaluación de las relaciones causales correspondientes, se obtienen los valores que se utilizarán como inputs de las siguientes relaciones causales, según el orden de propagación establecido por las conexiones del DAG G .

El resultado de la evaluación de la red bayesiana queda recogido en la variable o variables sumidero de la misma. A partir de los valores para todas las variables, incluida las variables sumidero, se calcula la distribución de probabilidad conjunta P de la red bayesiana, se valida el modelo con datos reales mediante el cálculo del ECM y se cuantifica la incertidumbre del modelo mediante el cálculo del índice de Shannon. El proceso de ajuste de la función P se lleva a cabo de forma iterativa mediante la realización de intervenciones tales como la modificación de la estructura de las variables aleatorias, la alteración de las relaciones causales identificadas o la variación de los parámetros del procedimiento de cálculo. En todos los casos se evalúa tanto el error cometido en la estimación de variables de las que se disponga de datos reales, mediante el cálculo del ECM, como la reducción de la incertidumbre, provocada por el incremento de información útil, mediante el índice de Shannon. De esta forma, se desarrolla una teoría T sobre la incertidumbre que es plenamente operacional y al mismo tiempo permite mejorar, mediante la generación de información útil, la base de conocimiento sobre el fenómeno de interés.

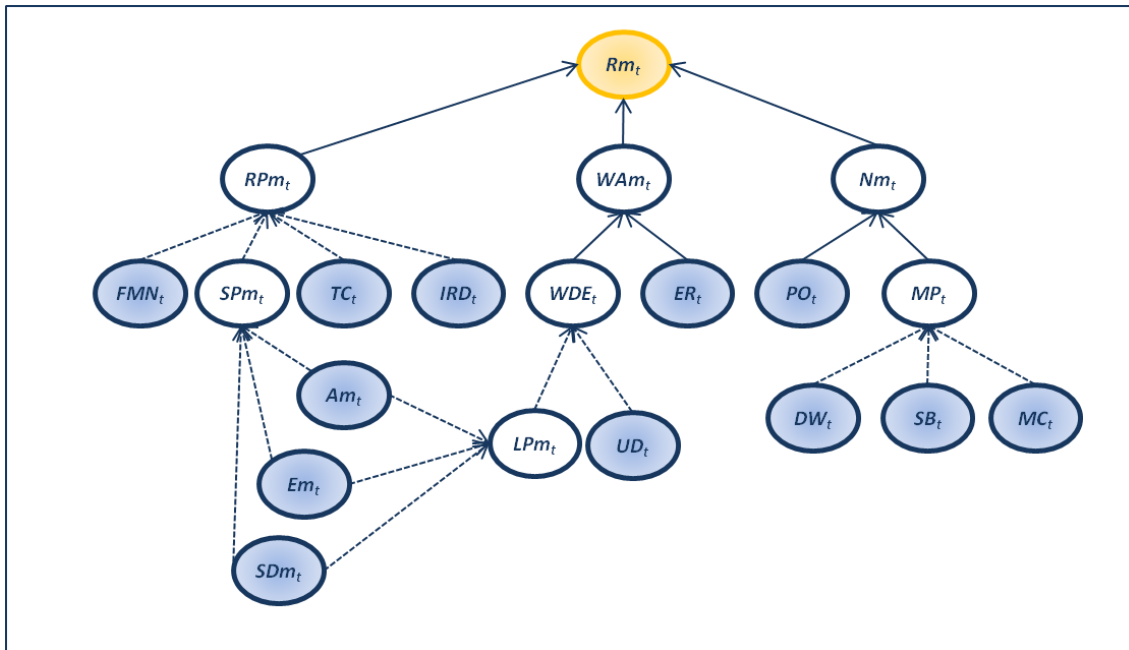
IV. RESULTADOS

En este capítulo se ofrecen los resultados de dos estudios complementarios sobre remesas y migraciones en los que se ha aplicado la teoría de la incertidumbre T desarrollada en la propuesta metodológica (Klir, 2006). En el primero de estos estudios se ha aproximado el flujo de remesas mediante el uso del procedimiento de cálculo simplificado. Así pues, una vez diseñada la red bayesiana causal sobre los flujos de remesas, esta es evaluada mediante el MSMC diseñado e hibridado con el MID-RB. En el segundo estudio, se aproxima el tamaño de la población migrante ecuatoriana en España, empleando para ello la sección *d-separada* de la red bayesiana y aplicando el procedimiento de cálculo mejorado con la inclusión del AEMO-RB, que permite estimar los pesos de las variables.

4.1. Estudio 1. Aplicación del procedimiento híbrido simplificado a la aproximación de los flujos de remesas en el corredor España–Ecuador

4.1.1. La red bayesiana causal para la estimación de los flujos de remesas

La red bayesiana, que representa el modelo conceptual sobre el fenómeno de las remesas, integra las variables explicativas de dichos flujos y las relaciones causales entre las mismas, identificadas mediante la revisión de la literatura científica ya mencionada en el Apartado 3.2.3. El DAG G , mostrado en la Figura 8, es compatible con la distribución de probabilidad conjunta P , puesto que se cumplen las dos condiciones de Markov. Así pues, cada una de las variables aleatorias X_i de la red (nodos), condicionada a sus padres markovianos PA_i , es independiente tanto de todos sus predecesores como de sus no descendientes, en el orden establecido mediante las flechas en G . Por tanto, es posible comprobar que, para los distintos caminos establecidos en la red bayesiana, existe un conjunto de nodos que bloquea los flujos de información entre variables, de tal forma que la probabilidad conjunta P de la red bayesiana puede ser potencialmente factorizada conforme a (Ec. 14).

Figura 8. Red bayesiana causal para la estimación de los flujos de remesas

Fuente: Elaboración propia.

Cabe señalar que G es una de las posibles aproximaciones a la representación de la probabilidad conjunta real P del fenómeno objeto de estudio, puesto que es posible desarrollar otros DAG o ampliar el diseñado en esta investigación mediante la inclusión de nuevas variables o nuevas relaciones causales relevantes que reflejen los avances en el estado del conocimiento sobre el fenómeno de las remesas.

En cuanto a la red bayesiana propuesta, esta se compone de veinte variables aleatorias continuas y ocho relaciones causales. En el DAG G (Figura 8) los nodos sombreados denotan las variables raíz de la red, mientras que el nodo destacado en color amarillo, que representa la variable relativa a los flujos de remesas Rm_t , es el sumidero de la red. Las relaciones causales de carácter algebraico han sido señaladas mediante flechas unidireccionales de trazo continuo, mientras que en las relaciones causales de carácter difuso el trazo de la flecha es discontinuo.

La Tabla 2 recoge de forma resumida estas relaciones. Las tres relaciones tipo algebraico han sido denotadas con f , mientras que las ocho de tipo difuso se han

denotado con u , referenciando la numeración asignada en el Apartado 3.2.3 de la propuesta metodológica donde han sido definidas.

Tabla 2. Relaciones causales de la RB

$Rm_t = f(RPm_t, WAm_t, Nm_t) = RPm_t \times WAm_t \times Nm_t$	(Ec. 6)
$RPm_t = u(FMN_t, TC_t, IRD_t, SP_t)$	(Ec. 7)
$SPm_t = u(Am_t, Em_t, SDm_t)$	(Ec. 8)
$WAm_t = f(WDE_t, ER_t) = WDE_t \times ER_t$	(Ec. 9)
$WDE_t = u(UD_t, LPm_t)$	(Ec. 10)
$LPm_t = u(Am_t, Em_t, SDm_t)$	(Ec. 11)
$Nm_t = f(PO_t, MP_t) = PO_t \times MP_t$	(Ec. 12)
$MP_t = u(SB_t, DW_t, MC_t)$	(Ec. 13)

Fuente: Elaboración propia.

4.1.2. Diseño experimental

4.1.2.1. Caso de estudio

La propuesta metodológica diseñada ha sido aplicada al caso de los flujos de remesas en el corredor España-Ecuador, siendo el objetivo determinar la validez de la red bayesiana propuesta mediante la estimación de los flujos de remesas en dicho corredor durante el periodo 2000-2010. Así pues, los resultados obtenidos mediante el modelo conceptual propuesto y la aplicación de la metodología diseñada son validados con la información suministrada por el Banco Central de Ecuador sobre los flujos de

remesas recibidos en el país durante dicho periodo⁶.

El corredor España-Ecuador es el décimo corredor más importante en el área de América Latina y el Caribe, con uno de los crecimientos más importantes tanto en número de inmigrantes como en el montante de los flujos monetarios remitidos (Ratha *et al.*, 2010). En los nueve corredores que lo preceden, el origen de los flujos financieros se encuentra en los Estados Unidos de América, ya que la proximidad geográfica, la estructura de su mercado laboral y la existencia de comunidades de migrantes bien asentadas, han convertido a este país en el destino tradicional de la población latinoamericana desde la década de los 70 (Bertoli *et al.*, 2011).

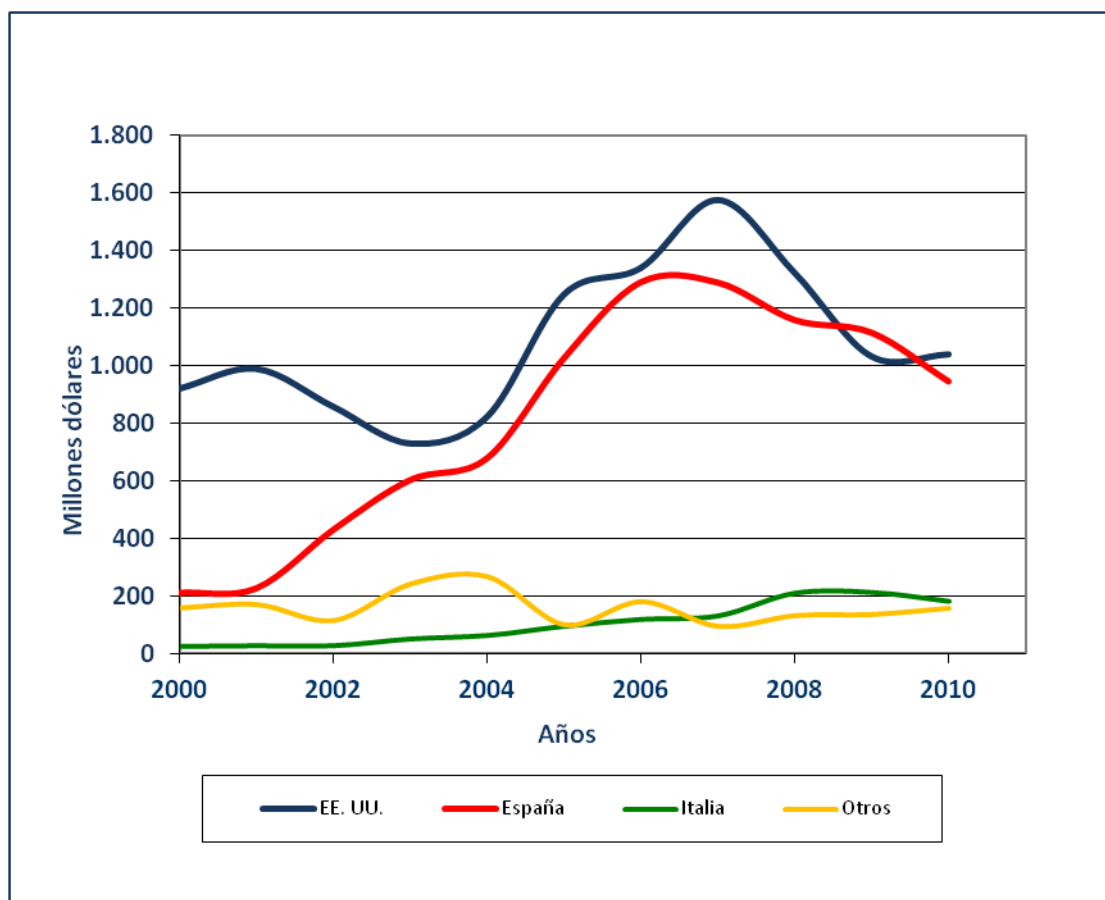
Ecuador cuenta con más de 12 millones de habitantes. De estos, un 60% y 30%, de los que habitan en zonas rurales y urbanas, respectivamente, vive en condiciones de pobreza (INEC, 2010), con un ingreso per cápita de 1.000 dólares o menos al año. La crisis económica e institucional que convulsionó el país durante los últimos años de la década de los noventa y principios de la década siguiente (Larrea, 2004), ha provocado un importante movimiento migratorio, situándose en 2010 la cifra total de emigrantes ecuatorianos en más de 1,1 millones de personas –el 8,3% de su población– (Ratha *et al.*, 2010).

Durante el periodo de estudio, España ha sido el país receptor de migrantes más importante de la UE, de tal forma que en 2010 el 12,3% de su población era inmigrante, frente al 6,6% de media en el conjunto de la UE (Vasileva, 2012). En relación a su stock de población inmigrante, los ecuatorianos son la cohorte, procedente del área de América Latina y el Caribe, más numerosa (399.000 personas), tras la población de origen rumano (831.000 personas) y el colectivo marroquí (754.000 personas) (INE, 2010). De esta forma, España se ha convertido en el primer

⁶ La disponibilidad de datos reales sobre otras variables también permitiría realizar las validaciones intermedias.

destino de la población emigrante ecuatoriana (Ratha *et al.*, 2010). Este hecho es más llamativo si se tiene en cuenta que la llegada de inmigrantes ecuatorianos a España se ha producido durante la última década (FLACSO, 2006), impulsados por una mayor flexibilidad en las condiciones de entrada en España en relación a Estados Unidos, destino tradicional de los emigrantes ecuatorianos (Bertoli *et al.*, 2011). Así, en el año 2010, España recibió al 35% de los emigrantes de esta nacionalidad (INEC, 2010).

Figura 9. Evolución de las remesas recibidas en Ecuador por país de origen



Fuente: Banco Central de Ecuador(2012).

El carácter económico de los flujos migratorios entre Ecuador y España se pone de relieve al observar la evolución del flujo de remesas con destino al país andino. Estos flujos alcanzaron su máximo en el periodo 2006-2007, como se puede observar en la Figura 9, cuando superaron los 1.280 millones de dólares anuales. Sin embargo, y acorde con la tendencia global, la crisis financiera ha provocado un descenso

acumulado hasta el año 2010 del 18% en las cantidades remitidas desde España (Banco Central de Ecuador, 2012). Aún así, y junto a Estados Unidos, España sigue siendo una de las principales fuentes de remesas de Ecuador.

De este modo, y a pesar de la crisis económica y de la contracción de estos flujos, las remesas enviadas por los inmigrantes ecuatorianos residentes en España durante el año 2010 ascendieron a 945 MM de dólares, casi el 41% del total recibido y equivalente al 1,6% del PIB de Ecuador (Banco Central de Ecuador, 2012).

4.1.2.2. Parámetros del procedimiento de cálculo

4.1.2.2.1. Estructura de las variables aleatorias

A partir de la información disponible, ha sido posible determinar la estructura de las variables aleatorias de la red diseñada, así como los parámetros necesarios para la evaluación de las relaciones causales difusas.

La Tabla 3 refleja las características de las variables raíz de la red bayesiana, que describen la situación socio-económica de Ecuador en comparación con la de España durante el periodo de estudio (2000-2010). El parámetro *evolución temporal* permite indicar la evolución de los valores de la variable a lo largo del periodo de estudio. Este parámetro puede tomar tres valores como son creciente, decreciente y fluctuante. Como su nombre indica, en los dos primeros casos, los datos de la variable muestran una evolución creciente/decreciente, mientras que en el tercero los valores se mueven o fluctúan en torno a un valor medio.

Tabla 3. Características de los inputs de la RB

Variable	Descripción	Unidad	Evolución temporal	Distribución estadística ^(a)	Fuente
Am_t	Edad	Años	Creciente	T[22, 27, 35]	INE (2010)
E_{mt}	Nivel educativo	Escala	Creciente	U [40, 50]	INE (2009); INEC (2008)
SDm_t	Duración de la estancia	Años	Creciente	U [0, 10]	INE (2009); Reher Sullivan <i>et al.</i> (2009)
FMN_t	Necesidades económicas familiares	Porcentaje	Decreciente	U [0,05, 0,45]	INEC (2011a)
TC_t	Costes de envío	Porcentaje	Fluctuante	T [0,02, 0,06, 0,09]	Remesas.org (2011); Banco Mundial (2011a)
IRD_t	Diferencial de tipo de interés	Porcentaje	Creciente	U [0,01, 0,05]	Banco Central de Ecuador(2011); Banco de España (2011)
UD_t	Tasa de desempleo	Porcentaje	Creciente	U [0,09, 0,30]	INE (2011a)
ER_t	Tipo de cambio	Euro/Dólar	Creciente	U [0,87, 1,55]	Banco de España (2011)
DW_t	Diferencial de salarios	Ratio	Decreciente	U [3,70, 4,90]	Banco Mundial (2011b)
SB_t	Beneficios sociales	Dólares per cápita	Creciente	U [1.850, 3.600]	OCDE (2011)
MC_t	Costes de la emigración	Dólares per cápita	Decreciente	U [1.500, 5.000]	Jokisch y Pribilsky (2002)
PO_t	Población en origen	Habitantes (millones)	Creciente	U [7,25, 8,8]	INEC (2011b)

^(a) Distribución estadística: triangular (T), uniforme (U).

Fuente: Elaboración propia.

La distribución estadística, que describe el comportamiento de la variable correspondiente a lo largo del periodo de estudio, ha sido determinada a partir de los datos disponibles en las fuentes indicadas en la tabla. Los valores extremos de la distribución se corresponden con los valores iniciales y finales de la variable durante dicho periodo. En la mayoría de los casos se ha asignado una distribución de tipo uniforme, denotada en la tabla como U, puesto que no hay indicios de que exista un valor más probable para la variable. En aquellos casos en los que este valor ha sido

identificado, la distribución asignada ha sido de tipo triangular, denotada como T. Cabe señalar que, la estructura de las variables aleatorias mostrada en esta tabla, es la mejor configuración posible, teniendo en cuenta los datos disponibles y los resultados obtenidos tras realizar numerosos experimentos con el modelo.

La información recogida por los Institutos de Estadística de Ecuador (INEC) y España (INE) a través de diversas encuestas ha permitido caracterizar la población migrante ecuatoriana residente en España. Por un lado, se ha empleado el informe elaborado por el INEC (2008) que caracteriza la emigración ecuatoriana a través de cuatro fuentes estadísticas, como son el Censo de Población y Vivienda correspondiente al año 2001, la Encuesta de Condiciones de Vida realizada durante el periodo 2005-2006 y las Encuestas de Empleo Desempleo y Subempleo de los años 2006 y 2007, ofreciendo información general sobre el perfil socio-demográfico del emigrante ecuatoriano, así como alguna información relativa a aquellos cuyo destino migratorio era España. Por otra parte, se ha utilizado la Encuesta sobre Inmigración realizada por el INE (2008) durante el periodo 2006-2007, junto con el documento de trabajo que acompaña a dicha encuesta (INE, 2009) y la monografía dedicada a la inmigración ecuatoriana realizada por Reher Sullivan *et al.* (2009), así como la explotación estadística del Padrón de Habitantes (INE, 2010).

Si bien no se dispone de datos a lo largo de todo el periodo, salvo para el caso de la edad, el uso de estas fuentes ha permitido caracterizar a los ecuatorianos residentes en España como una población joven, ya que la mayoría tienen edades Am_t comprendidas entre los 22 y 35 años, aunque la edad media suele estar en torno a los 27 años; con un nivel educativo medio-bajo Em_t , puesto que el porcentaje con estudios superiores supone menos de una décima parte de la población inmigrante ecuatoriana, más de la mitad han completado la educación secundaria y el resto sólo ha cursado estudios básicos; y una estancia media SDm_t en el país entre los 8 y 9 años. En los dos últimos casos, la evolución temporal se ha supuesto creciente puesto que proceso migratorio ha sido protagonizado por las clases medias ecuatorianas, suponiendo para el país una cierta “fuga de cerebros”, al tiempo que los inmigrantes

ecuatorianos manifiestan, en su mayoría, la intención de quedarse en España (Reher Sullivan *et al.*, 2009).

Las otras tres variables raíz que intervienen en la determinación de la propensión a remitir del migrante han sido determinadas a partir de las series temporales facilitadas por distintos organismos. Así pues, las necesidades económicas de la familia del migrante FMN_t han sido aproximadas mediante el indicador de restricción al consumo (INEC, 2011a). Este indicador refleja en que porcentaje el coste de una cesta de la compra básica no puede ser cubierto con los ingresos por rentas del trabajo de un hogar medio ecuatoriano⁷. Los costes de transacción asociados al envío de remesas al país de origen TC_t , han sido fijados a partir del coste medio de envío de una remesa de 150 euros desde España a Ecuador, a partir de las encuestas periódicas realizadas por la organización Remesas.org (2011) y de los datos sobre precios de remesas compilados por el Banco Mundial (2011a), si bien en ambos casos los datos solo están disponibles desde el año 2005. La tercera variable, el diferencial de tipo de interés entre el país de origen y el país de residencia del migrante IRD_t , ha sido calculada a partir de las series temporales accesibles en las webs del Banco Central de Ecuador y del Banco de España. Para ello se han comparado las tasas de interés efectivas medias para depósitos sin distinción alguna por plazo o tipo de agente que realiza el depósito. En el caso de España estas series están disponibles para todo el periodo, mientras que en el caso de Ecuador, solo se disponen de series completas a partir de 2007, como consecuencia de la implementación de cambios metodológicos. En este último caso, la información sobre los periodos anteriores ha sido completada mediante publicaciones monográficas del Banco Central de Ecuador (2010).

⁷ La canasta familiar básica fijada por el INEC está compuesta por 75 de los 299 productos que conforma la cesta de bienes del IPC para Ecuador, reflejando el gasto mensual que debe realizar una familia (4 miembros) para la adquisición de bienes y servicios imprescindibles para satisfacer las necesidades básicas del hogar (alimentos y bebidas, vivienda, indumentaria y misceláneos). Los ingresos percibidos por el hogar se calculan a partir del salario básico unificado fijado anualmente por el Gobierno de Ecuador, considerando 1,6 perceptores del mismo (INEC, 2011a).

Para la mayoría de las restantes variables raíz se disponen de series completas para el periodo de estudio. Así, la evolución de la tasa de desempleo UD_t ha sido aproximada a partir de los datos de la Encuesta de Población Activa (INE, 2011a), concretamente mediante la tasa de desempleo para extranjeros no pertenecientes a la Unión Europea. La serie para el tipo de cambio ER_t del euro respecto al dólar ha sido obtenida del Banco de España (2011), mientras que el diferencial de salarios DW_t entre España y Ecuador ha sido aproximado mediante la comparación del Producto Interior Bruto en paridad de poder adquisitivo (Jennissen, 2003) para ambos países, cuyos datos han sido obtenidos del Banco Mundial (2011b). Los datos sobre la población ecuatoriana PO_t , que hacen referencia a la evolución de la población en edad laboral - entre 18 y 64 años, han sido obtenidos de la proyección de población realizada por el INEC (2011b) para el periodo 2000-2010.

En cuanto a la cuantificación de los beneficios sociales SB_t percibidos por los inmigrantes en España, esta se ha llevado a cabo mediante la base de datos compilada por la OCDE sobre gasto social (OCDE, 2011) que contaba con registros hasta el año 2007. Solo se han considerado aquellos programas de gastos relacionados con salud, educación y políticas activas de empleo, así como otros programas residuales de subsidios relativos a la atención a personas mayores y discapacidad, programas de vivienda, servicios a familia, etc. No se han incluido las partidas de pensiones ni desempleo, ya que la situación laboral de los inmigrantes impide, con frecuencia, que generen derechos sobre este tipo de prestaciones sociales (Moreno y Bruquetas, 2011). Finalmente, para aproximar los costes monetarios de la migración MC_t sólo se dispone de la información recogida en los estudios de Jokisch y Pribilsky (2002). Estos costes solo hacen referencia a los costes asociados al viaje, por lo que han sido incrementados teniendo en cuenta los costes administrativos y de subsistencia del inmigrante hasta que encuentra su primer empleo en España. Además se ha supuesto que dicho coste ha ido disminuyendo a medida que las redes sociales de migrantes ecuatorianos han ido adquiriendo mayor tamaño.

Por otra parte, cabe señalar que el MSMC permite modelar el comportamiento no lineal de las variables aleatorias a lo largo del tiempo, para ello el periodo temporal de referencia se divide en intervalos y se indica la variación que experimenta la pendiente de la curva (tanto por uno), respecto a la pendiente correspondiente al intervalo de tiempo precedente. Los porcentajes asignados se muestran en la Tabla 4. Las gráficas correspondientes a los valores reales, en aquellos casos en los que se disponga de los mismos, y a los valores estimados obtenidos se muestran en el apartado correspondiente a la estimación de los flujos de remesas (Apartado 4.1.4).

Tabla 4. Variación de la pendiente de la curva para las variables del modelo con comportamiento no lineal durante el periodo de estudio

Intervalos tiempo	Variables							
	SDm_t	FMN_t	UD_t	ER_t	SB_t	DW_t	PO_t	MC_t
2000-2001	1,2	-0,56	0,18	-0,51	0,15	-0,06	0,29	-5
2001-2002		0,12	0	0,15	0,22	0,01	0,47	
2002-2003		-0,1	0,05	0,56	0,31	-0,03		
2003-2004		-0,04	-0,13	0,34	0,21	-0,44		
2004-2005		0,19	-0,5	0	0,28	-0,17		
2005-2006			-0,06	0,02	0,46	-0,08	0,3	
2006-2007	0,2	-0,19	0,11	1	0,5	0,11	0,5	0,15
2007-2008			0,34			-0,48		0,03
2008-2009		-0,34	0,68		0,05	-0,34		
2009-2010		-0,2	0,12	-0,38	-0,19			
2010-2011		-0,27						

Fuente: Elaboración propia.

4.1.2.2.2. Parámetros para la evaluación de relaciones difusas

El otro bloque de información necesario para poder evaluar las relaciones causales del modelo mediante el procedimiento de cálculo propuesto lo constituyen los parámetros que permiten el diseño de las reglas difusas. Esta información se muestra en dos tablas. La primera de ellas, Tabla 5, recoge la información relativa a aquellas variables que son inputs en una relación difusa, mientras que en la segunda, Tabla 6, se muestra la información sobre los outputs de estas relaciones. El orden establecido por las flechas del DAG G determina que algunas variables sean output en una relación difusa e inputs de la siguiente, por lo que se ha optado por incluirlas en ambas tablas, indicando en cada caso los parámetros correspondientes.

La Tabla 5 recoge la información requerida sobre los inputs de las Relaciones Difusas (RD), es decir, el Universo de Discurso de la variable input (UD), el tipo de influencia o de relación causal entre el input y el output en cada relación, y el peso o nivel de relevancia de cada input sobre el correspondiente output.

Según G , las variables raíz del modelo son, a su vez, variables inputs de las relaciones difusas, de tal forma que el UD de estas variables coincide con el intervalo de valores de las correspondientes distribuciones estadísticas. Sólo dos variables, el perfil socio-demográfico SPm_t y el perfil laboral LPM_t del inmigrante, no fueron definidas en la Tabla 3, puesto que se trata de variables con esa doble función output/input. Dado que no existe información sobre estas variables en la literatura, se ha establecido como UD todo el intervalo de posibles valores de la unidad de medida seleccionada. En cuanto al tipo de relación causal, esta puede ser de tipo positivo o negativo, si la relación entre la variable input y la variable output detectada en la literatura es, respectivamente, directa o inversa.

Tabla 5. Relaciones de dependencia difusa: características de las variables inputs

Variable	Descripción	Unidad	RD	UD	Influencia	Pesos (w_i) ^(a)
Am_t	Edad	Años	8	[22, 35]	Negativa	T[0.09, 0.10, 0.11]
E_{mt}	Nivel educativo	Escala	8	[40, 50]	Negativa	T[0.09, 0.10, 0.11]
SDm_t	Duración de la estancia	Años	8	[0, 10]	Positiva	T[0.79, 0.80, 0.81]
Am_t	Edad	Años	11	[22, 35]	Positiva	T[0.29, 0.30, 0.31]
E_{mt}	Nivel educativo	Escala	11	[40, 50]	Positiva	T[0.09, 0.10, 0.11]
SDm_t	Duración de la estancia	Años	11	[0, 10]	Positiva	T[0.59, 0.60, 0.61]
SPm_t	Perfil socio-demográfico	Escala	7	[0, 100]	Positiva	T[0.54, 0.55, 0.56]
FMN_t	Necesidades económicas familiares	Porcentaje	7	[0.05, 0.45]	Positiva	T[0.29, 0.30, 0.31]
TC_t	Costes de envío	Porcentaje	7	[0.02, 0.09]	Negativa	T[0.09, 0.10, 0.11]
IRD_t	Diferencial de tipo de interés	Porcentaje	7	[0.01, 0.05]	Positiva	T[0.04, 0.05, 0.06]
LPM_t	Perfil laboral	Escala	10	[0, 100]	Positiva	T[0.39, 0.40, 0.41]
UD_t	Tasa de desempleo	Porcentaje	10	[0.09, 0.30]	Negativa	T[0.59, 0.60, 0.61]
DW_t	Diferencial de salarios	Ratio	13	[3.70, 4.90]	Positiva	T[0.14, 0.15, 0.16]
SB_t	Beneficios sociales	Dólares per cápita	13	[1850, 3600]	Positiva	T[0.04, 0.05, 0.06]
MC_t	Costes de la emigración	Dólares per cápita	13	[1500, 5000]	Negativa	T[0.79, 0.80, 0.81]

^(a) Distribución estadística triangular (T)

Fuente: Elaboración propia.

En aquellos casos en los que la literatura empírica al respecto era contradictoria, es decir, existían evidencias de relaciones tanto de tipo directo como inverso, se ha recurrido a dos vías para caracterizar dichas relaciones. Por un lado, se han empleado artículos relativos al colectivo migrante ecuatoriano en España (Campoy-Muñoz *et al.*, 2013; Echazarra, 2010). Por otro lado, se han realizado sucesivas intervenciones en la red bayesiana, con objeto de lograr un mejor ajuste de la distribución de probabilidad conjunta P , siendo éstas evaluadas mediante el ECM y

la Entropía de Shannon. De tal forma, que mediante la primera vía se ha establecido una influencia negativa para el nivel de estudios, y mediante la segunda se han caracterizado el resto de relaciones relativas al perfil socio-demográfico del migrante, tal y como se muestra en la Tabla 5.

El proceso de ajuste de la función P se lleva a cabo de forma iterativa mediante la realización de intervenciones tales como la modificación de la estructura de las variables aleatorias, la alteración de las relaciones causales identificadas o la variación de los parámetros del procedimiento de cálculo. En todos los casos se evalúa tanto el error cometido en la estimación de variables de las que se dispone de datos reales, mediante el cálculo del ECM, como la reducción de la incertidumbre, provocada por el incremento de información útil, mediante el índice de Shannon. Esta tarea resulta tediosa, especialmente en modelos complejos como el que se presenta en esta aplicación, lo que justifica, entre otras causas, la implementación del AEMO-RB.

En cuanto a la determinación de los pesos w_i , la distribución estadística asignada ha sido determinada, de nuevo, en base tanto a la literatura empírica como a la experimentación. En algunos estudios, la literatura empírica ha permitido establecer un orden de relevancia entre las variables. Así pues, el hecho de que un gran porcentaje de las remesas se destine a gastos de consumo (INEC, 2008), así como la importancia que han demostrado las características relativas al perfil socio-demográfico del migrante (Campoy-Muñoz *et al.*, 2013) ha determinado la asignación de pesos reflejada en la Tabla 5. Además, la naturaleza dual del mercado laboral español, en el que los inmigrantes suelen ocuparse en empleos con las peores condiciones laborales y la mayor sensibilidad de este grupo a los shocks económicos (Simón *et al.*, 2008), ha determinado que la tasa de desempleo tenga mayor relevancia que el perfil laboral del migrante de cara a la determinación del salario. Finalmente, la propensión a migrar a España de la población ecuatoriana se ve principalmente influenciada por un menor coste de la migración en relación a Estados Unidos, destino tradicional de la población que decidía emigrar, mientras que el diferencial de salarios tiene menos relevancia Bertoli *et al.* (2011).

En la Tabla 6, se muestran los parámetros relativos a los outputs de las relaciones difusas. Cabe señalar que, dado el estado de conocimiento actual sobre el corredor analizado, el *tipo* e *intensidad* de las relaciones difusas ha sido fijado como neutras, esto significa que los parámetros Q y D tomarán valores cero en la evaluación de las relaciones correspondientes. Por otra parte, se ha introducido una perturbación aleatoria en el modelo, de forma que las relaciones difusas son instanciadas y evaluadas con una probabilidad (Prob.) de 0,98 por el MID-RB, mientras que en las restantes ocasiones el valor del output es determinado por el MSMC mediante una distribución uniforme definida sobre el UD identificado para estas variables.

Tabla 6. Relaciones de dependencia difusas: características de las variables outputs

Variable	Descripción	Unidad	RD	UD	Tipo	Intensidad	Prob	Fuente
RPm_t	Propensión a remitir	Porcentaje	7	[0.10, 0.35]	Neutra	Neutral	0,98	Moré <i>et al.</i> (2008)
SPm_t	Perfil socio-demográfico	Escala	8	[0, 100]	Neutra	Neutral	0,98	-
WDE_t	Salario en el país de residencia	Euros	10	[6000, 11000]	Neutra	Neutral	0,98	(INE, 2011b)
LPm_t	Perfil laboral	Escala	11	[0, 100]	Neutra	Neutral	0,98	-
MP_t	Propensión a emigrar	Porcentaje	13	[0.01, 0.05]	Neutra	Neutral	0,98	INEC (2011b, 2010)

Fuente: Elaboración propia.

El UD de las variables outputs es determinado en base a la información disponible, salvo en el caso de SPm_t y LPm_t , como ya se ha comentado. En el caso de la propensión a remitir, el estudio de Moré *et al.* (2008) determina la propensión a remitir para los emigrantes ecuatorianos en España para 2006. En cuanto a los salarios de los inmigrantes ecuatorianos, este intervalo ha sido establecido a partir de la información recogida en la Encuesta de Estructura Salarial realizada por el INE (2011b). Este salario es una media ponderada del salario según las categorías establecidas en la Clasificación Nacional de Ocupaciones. Estas ponderaciones han sido establecidas según la distribución de los trabajadores ecuatorianos por categoría profesional

recogida en Reher Sullivan *et al.* (2009). Adicionalmente, el salario medio obtenido, ha sido minorado en 20% ya que este es el porcentaje medio de diferencia entre salarios trabajadores inmigrantes y nativos en España (Izquierdo *et al.*, 2009). La propensión a emigrar a España de la población ecuatoriana en edad laboral ha sido determinado como un ratio entre las salidas registradas de ecuatorianos con destino a España, registradas en el Sistema de Entradas y Salidas Internacionales (INEC, 2010), y la población ecuatoriana en edad laboral recogida en las Proyecciones de Población (INEC, 2011b).

Por último, se ha establecido que el número de conjuntos difusos tanto para las variables inputs como outputs de las relaciones difusas es de $e=31$ funciones de pertenencia para cada uno de los $n=15$ inputs de las $h=5$ relaciones difusas, por lo que el sistema puede llegar a evaluar un número de reglas de $1.17 \cdot 10^{23}$ reglas (Ec. 25). Las funciones de pertenencia empleadas son de tipo triangular, dado su simplicidad y solidas bases teóricas (Pedrycz, 1994), y sigmoideal para los extremos, con solapamiento perfecto dos a dos. El número de simulaciones fijadas para el MSCM ha sido de $m=11.000$ divididas en $b=10$ lotes, con eliminación de las 100 primeras simulaciones de cada lote.

4.1.3. Caracterización del proceso de simulación

La Tabla 7 recoge el análisis del proceso de simulación mediante el cálculo del error según la propuesta de Nakayama (Ec. 21). La variabilidad del proceso de simulación se caracteriza como el porcentaje que representa el error sobre el valor medio estimado de cada una de las variables a lo largo del periodo temporal considerado, indicando el valor medio, así como el valor mínimo y máximo del error para cada una de las variables. Esta medida apenas supera el 1.5% en el peor de los casos, lo cual es indicativo de que el proceso de simulación es muy estable, es decir, a medida que el número de simulaciones aumenta, los valores obtenidos convergen hacia un valor central del intervalo de confianza definido.

Tabla 7. Descriptivos sobre la medida de variabilidad del MSMC

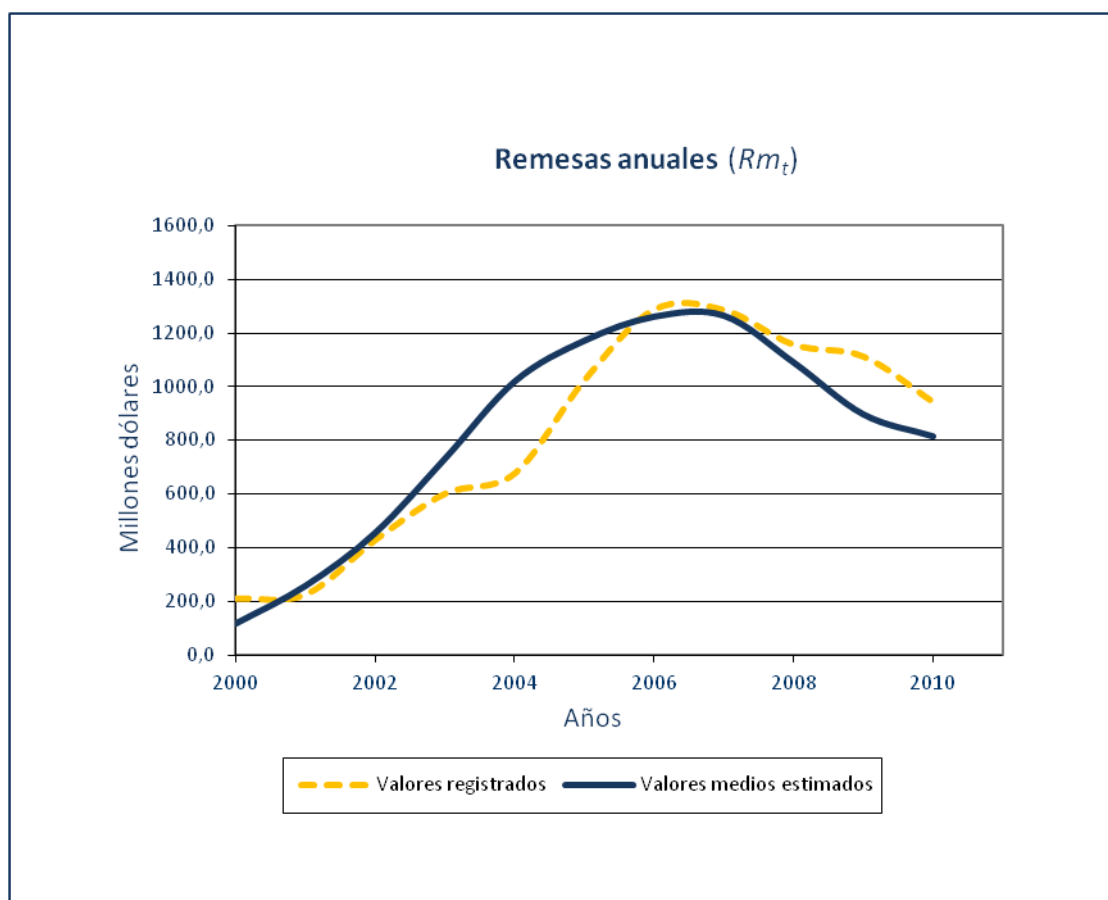
Variable	Descripción	Error de Nakayama (%)		
		Media	Mínimo	Máximo
Em_t	Nivel educativo	0,05	0,04	0,06
Am_t	Edad	0,07	0,06	0,08
SDm_t	Duración de la estancia	0,46	0,09	1,20
SPm_t	Perfil socio-demográfico	0,30	0,11	0,67
FMN_t	Necesidad económicas familiares	0,94	0,10	1,41
TC_t	Costes envío	0,53	0,34	0,84
IRD_t	Diferencial de tipo de interés	0,32	0,20	0,44
RPM_t	Propensión a remitir	0,18	0,11	0,31
LPm_t	Perfil laboral	0,43	0,12	1,23
UD_t	Tasa de desempleo	0,50	0,26	0,67
WAm_t	Salario en la moneda del país de origen del migrante	0,16	0,09	0,21
ER_t	Tipo de cambio	0,16	0,06	0,24
WDE_t	Salario en la moneda del país de residencia del migrante	0,22	0,13	0,28
DW_t	Diferencial de salarios	0,05	0,02	0,07
SB_t	Beneficios sociales en el país de residencia del migrante	0,22	0,12	0,34
MC_t	Costes de la emigración	0,54	0,06	0,86
MP_t	Propensión a emigrar	0,43	0,24	1,19
PO_t	Población en el país de origen del migrante	0,04	0,02	0,06
Nm_t	Stock de migrantes	0,44	0,27	1,18
Rm_t	Remesas	0,52	0,29	1,10

Fuente: Elaboración propia.

4.1.4. Estimaciones de flujos de remesas y validación con datos reales

La Figura 10, muestra los resultados obtenidos así como los flujos de remesas registrados por el Banco Central de Ecuador para el periodo de estudio (2000-2010).

Figura 10. Flujo de remesas en el corredor España-Ecuador durante el periodo 2000-2010 (millones de dólares): valores registrados (línea discontinua) versus valores estimados (línea continua)



Fuente: Elaboración propia.

Los valores estimados ajustan la tendencia global de los flujos de remesas registrados, alcanzando el máximo registrado durante en el periodo 2006-2007. A primera vista, el gráfico nos permite distinguir dos periodos, un primer periodo que abarcaría desde el inicio del estudio hasta el año 2005, y un segundo periodo, que se inicia a partir de 2006 y alcanza hasta el final de periodo de estudio. Durante el primer

periodo (2000-2005), los flujos de remesas estimadas superan los flujos registrados, que experimentan un sorprendente crecimiento, pasando de 210 millones de dólares a casi 1025 millones de dólares en apenas 6 años. Por el contrario, el modelo subestima los flujos de remesas que se producen durante el segundo periodo (2006-2010), mostrando una reducción mayor que la registrada (30%) como consecuencia de la crisis económica mundial (Banco Central de Ecuador, 2012). El Error Cuadrático Medio (ECM), que se muestra en la Tabla 8 permite evaluar las diferencias entre los flujos medios estimados y los flujos registrados de forma más precisa.

Tabla 8. Descriptivos para los flujos de remesas estimados en el corredor España-Ecuador (Millones de dólares)

Año	Valor registrado	Estimación media	Desviación estándar	ECM ^(a)	IC ^(b)
2000	210.67	120.47	1.35	97.63	[119.52, 121.41]
2001	226.32	260.85	5.21	146.86	[257.18, 264.53]
2002	429.60	457.71	5.61	193.32	[453.76, 461.66]
2003	602.15	733.65	8.95	285.23	[727.34, 739.96]
2004	677.86	1020.17	9.22	455.87	[1013.67, 1026.67]
2005	1025.23	1171.63	8.38	360.44	[1165.72, 1177.53]
2006	1288.87	1260.62	12.91	350.97	[1251.52, 1269.72]
2007	1287.40	1264.19	11.68	362.88	[1255.97, 1272.42]
2008	1157.50	1090.16	8.74	329.15	[1084.00, 1096.32]
2009	1113.40	897.43	8.06	338.60	[891.75, 903.10]
2010	944.40	815.88	7.64	273.23	[810.50, 821.26]

^(a) Error cuadrático medio: valores estimados *versus* valores registrados.

^(b) Intervalo de confianza ($\alpha=0.05$).

Fuente: Elaboración propia.

Considerando la totalidad del periodo 2000-2010, el ECM alcanza un valor medio de 290.18 millones dólares (desviación estándar: ± 106.09). Mientras que por

periodos, el ECM alcanza un valor medio de 256.56 millones de dólares (desviación estándar: ± 136.11) durante el primer periodo (2000-2005) y de 330.96 millones de dólares (desviación estándar: ± 34.96) durante el Segundo (2006-2010), aunque el ECM es menor durante el primer periodo, también presenta una mayor variabilidad.

A continuación se muestran los valores medios estimados para las distintas variables de la red bayesiana (Figuras 11 a 14). En la Figura 11 y en la Figura 12 se muestran los valores estimados para la sección de la red relativa a la propensión a remitir del inmigrante RPM_t . En la primera de estas figuras se muestra las variables antecedentes de dicha propensión, es decir, las necesidades económicas de las familias en Ecuador (FMN_t), el perfil socio-demográfico del migrante ecuatoriano residente en España (SPM_t), así como los costes de envío de remesas desde España a Ecuador (TC_t) y el diferencial de tipos de interés entre ambos países (IRD_t). En la segunda figura se muestran las variables que determinan el perfil socio-demográfico del migrante, la edad (Am_t), el nivel educativo (Em_t) y la duración de la estancia en España (SDM_t). Además, dado que estas son las mismas variables que determinan el perfil laboral del inmigrante (LPm_t), esta última variable también ha sido incluida en dicha figura.

En la Figura 13 se recogen los valores estimados correspondientes a la sección de la red relativa al salario de inmigrante. Así pues se muestra el salario del inmigrante en moneda de curso legal en su país de origen (WAm_t), que en el caso de Ecuador es el dólar, y en la moneda del país de residencia, que en el caso de España se trata del euro, así como la evolución del tipo de cambio (ER_t) y de la tasa de desempleo en España (UD_t). En esta figura se vuelve a incluir la gráfica correspondiente al perfil laboral del migrante (LPm_t), con objeto de ofrecer una visión de conjunto de esta sección de la red. Por último, en la Figura 14 se presentan los valores estimados para la sección de la red relativa a la población ecuatoriana en España. De esta forma, junto al stock de inmigrantes (Nm_t), se muestran los resultados obtenidos para la población ecuatoriana en edad laboral (POt) y las estimaciones sobre su propensión a emigrar a España España (MP_t), así como la evolución los distintos determinantes de dicha propensión, es decir, el diferencial de salarios entre España y Ecuador (DW_t), el

importe monetario de los beneficios sociales existentes en España (SB_t) y de los costes asociados a la migración (MC_t).

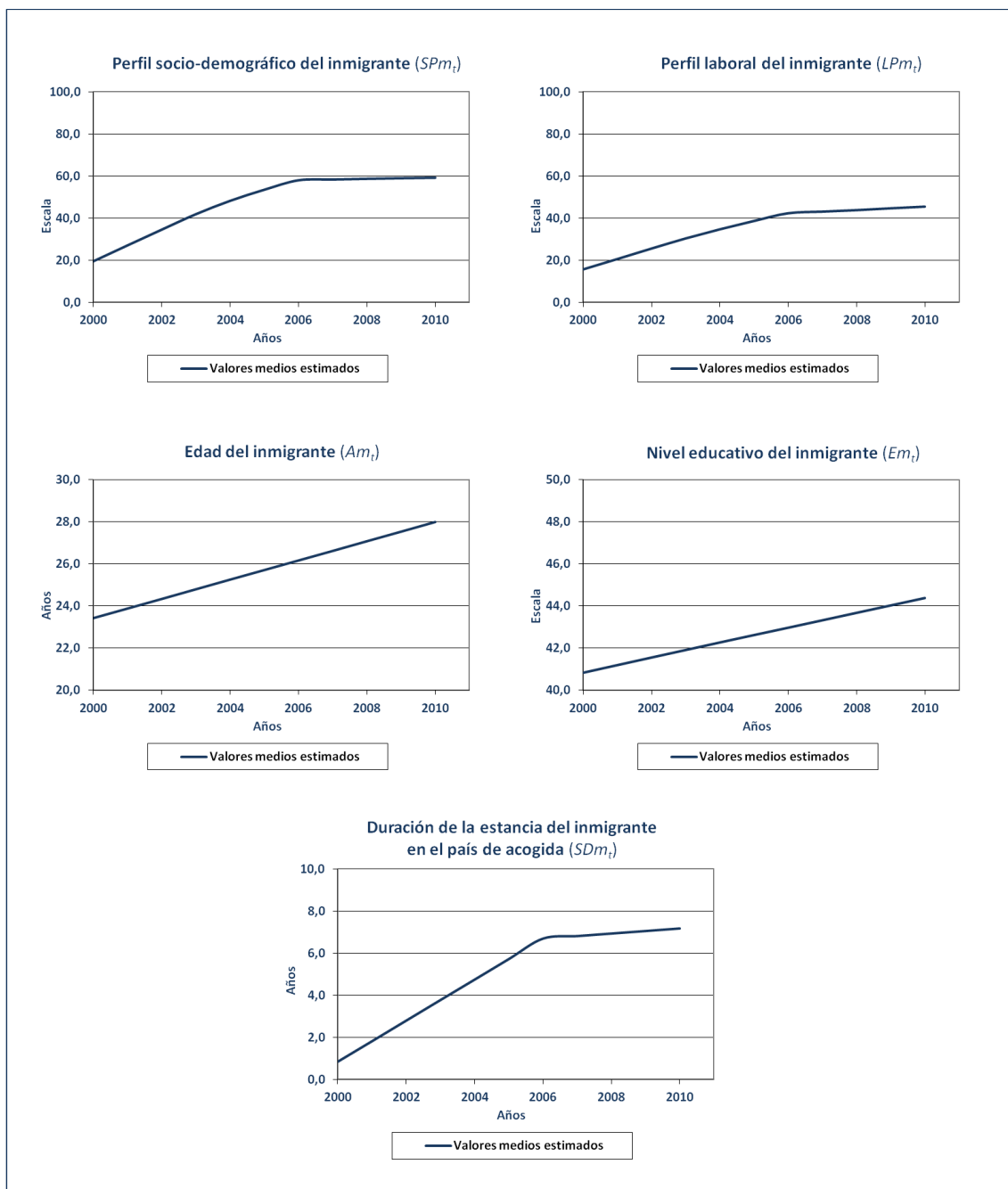
Figura 11. Valores estimados para las variables de la red relativas a la propensión media a remitir de los inmigrantes ecuatorianos en España^a



(a) Los datos reales de las variables son incluidos en la gráfica correspondiente con trazo discontinuo, cuando se encuentran disponibles.

Fuente: Elaboración propia.

Figura 12. Valores estimados para las variables de la red relativas al perfil socio-demográfico y laboral del inmigrante ecuatoriano en España^a



(a) Los datos reales de las variables son incluidos en la gráfica correspondiente con trazo discontinuo, cuando se encuentran disponibles.

Fuente: Elaboración propia.

Figura 13. Valores estimados para las variables de la red relativas al salario medio del inmigrante ecuatoriano en España^a



(a) Los datos reales de las variables son incluidos en la gráfica correspondiente con trazo discontinuo, cuando se encuentran disponibles.

Fuente: Elaboración propia.

Figura 14. Valores estimados para las variables de la red relativas al tamaño medio del stock de población inmigrante ecuatoriana en España^a



(a) Los datos reales de las variables son incluidos en la gráfica correspondiente con trazo discontinuo, cuando se encuentran disponibles.

Fuente: Elaboración propia.

4.1.5. Distribución de probabilidad conjunta y Entropía de Shannon

Para estimar la Entropía de Shannon de la red bayesiana, los flujos de remesas estimados han sido discretizados usando distintos números de intervalos I . En la Tabla 9 se muestran los valores de la Entropía de Shannon para dichos intervalos. Como puede observarse, el valor del índice aumenta a medida que la variable es discretizada usando mayor número de intervalos. Este incremento es resultado de la evolución creciente de los valores del logaritmo (Ec. 35) a medida que las probabilidades asociadas a los correspondientes intervalos disminuyen.

Tabla 9. Entropía de Shannon

Año	$I=10$	$I=25$	$I=50$
2000	1,29	2,20	3,14
2001	2,26	3,47	4,44
2002	2,43	3,62	4,69
2003	2,58	3,87	4,85
2004	2,64	3,94	4,94
2005	2,52	3,81	4,81
2006	2,64	3,93	4,93
2007	2,49	3,89	4,89
2008	2,71	4,01	5,01
2009	2,63	3,91	4,91
2010	2,48	3,76	4,76

Fuente: Elaboración propia.

En la Tabla 10 se muestra el valor de la probabilidad de los flujos medios de remesas estimados según los intervalos establecidos para el índice anterior.

Tabla 10. Probabilidad de los flujos de remesas estimados para el periodo de estudio

Año	Valor medio estimado	Intervalo (I=10)	Prob.	Intervalo (I=25)	Prob.	Intervalo (I=50)	Prob.
2000	120,47	114,19-172,42	0,46	102,5-125,84	0,40	114,19-125,84	0,19
2001	260,85	156,09-266,92	0,27	222,59-266,92	0,09	244,76-266,92	0,04
2002	457,71	354,99-500,83	0,27	413,33-471,66	0,11	442,50-471,66	0,05
2003	733,65	621,30-798,17	0,28	727,42-798,17	0,11	727,42-762,79	0,06
2004	1020,17	970,91-1.172,99	0,26	970,91-1.051,74	0,10	1.011,33-1.051,74	0,05
2005	1171,63	1.102,65-1.343,46	0,29	1.102,65-1.198,97	0,12	1.150,81-1.198,97	0,06
2006	1260,62	1.140,57-1.374,48	0,27	1.234,14-1.327,70	0,11	1.234,14-1.280,91	0,06
2007	1264,19	1.132,01-1.380,75	0,37	1.231,51-1.331,01	0,12	1.231,51-1.281,26	0,06
2008	1090,16	971,39-1.175,82	0,27	1.053,16-1.134,93	0,11	1.053,16-1.094,05	0,06
2009	897,43	816,83-993,57	0,28	887,53-958,22	0,11	887,53-922,87	0,06
2010	815,88	810,58-991,84	0,31	810,58-883,08	0,14	810,58-846,83	0,07

Fuente: Elaboración propia.

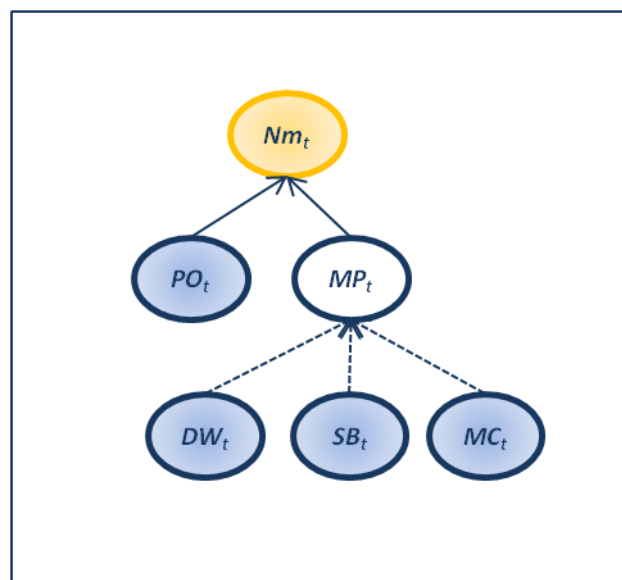
Cabe señalar que una disminución en el índice calculado es resultado de una disminución en la entropía, es decir, un aumento en la información útil generada por el modelo. Por tanto, dicho índice sirve de guía para la evaluación de las intervenciones realizadas en la red bayesiana, con objeto de conseguir un mejor ajuste de la función P , e incluso podría ser incluido como un nuevo objetivo en el AEMO-RB que conforma la propuesta metodológica. Sin embargo, el comportamiento mostrado por el índice pone de manifiesto una de sus principales debilidades, como es la determinación del número de intervalos idóneos para su discretización. Este hecho ha motivado que, finalmente, el índice no haya sido incluido en el diseño del AEMO-RB que se describe a continuación.

4.2. Estudio 2. Aplicación del procedimiento híbrido mejorado a la aproximación del tamaño de la población inmigrante ecuatoriana en España

4.2.1. La red bayesiana causal para la estimación del tamaño de la población inmigrante en el país de acogida

La aplicación del criterio de *d-separación* permite el estudio de secciones de una red bayesiana de forma autónoma. En base a esta propiedad, para este segundo estudio se ha elegido la sección *d-separada* que modela el tamaño o stock de población inmigrante en el país de acogida del migrante y, por tanto, origen de los flujos de remesas. En la Figura 15 se muestra dicha sección, mientras que en la Tabla 11 se recogen las correspondientes relaciones causales.

Figura 15. Sección *d-separada* de la red bayesiana causal para la estimación del tamaño de la población inmigrante en el país de acogida



Fuente: Elaboración propia.

Tabla 11. Relaciones causales de la sección *d-separada* de la red bayesiana

$$Nm_t = f(PO_t, MP_t) = PO_t \times MP_t \quad (Ec. 12)$$

$$MP_t = u(SB_t, DW_t, MC_t) \quad (Ec. 13)$$

Fuente: Elaboración propia.

La importancia del stock de migrantes Nm_t para la estimación de los flujos de remesas, puesto que determina el número de personas que potencialmente pueden enviar remesas, junto con la existencia de series de datos completas que estiman este stock, han sido los criterios que han guiado la elección de dicha sección de la red para la aplicación del procedimiento de cálculo mejorado.

4.2.2. Diseño experimental

4.2.2.1. Caso de estudio

Para comprobar el comportamiento del procedimiento de cálculo mejorado, en este segundo estudio se ha estimado el stock de población migrante ecuatoriana residente en España en el año $t=2000$. La elección de un único periodo viene motivada tanto por la complejidad computacional del proceso, como por la comparación con los pesos fijados en el Estudio 1. En el primer caso, dicha complejidad se debe a que, dada la naturaleza exploratoria del estudio, todos los parámetros de funcionamiento del AEMO-RB deben ser calibrados. En el segundo caso, en el Estudio 1 se establecen unos pesos para todo el periodo que, en el caso de la sección *d-separada* para el stock de migrantes, se han fundamentado en la relevancia de las variables al inicio del periodo de estudio por ser la única información disponible.

4.2.2.2. Parámetros del MSMC

Las Tabla 12 a 14 recogen la información relativa a los parámetros exigidos para

el MSCM. Estos parámetros son los mismos que en el Estudio 1, con la excepción del número de simulaciones y de lotes, que ha sido fijado en $m=500$ y $b=1$ respectivamente, para reducir el tiempo de computación. En este caso el MSCM debe evaluar una relación algebraica y otra difusa, por lo que el número de reglas puede ascender a 29.971 (Ec. 25) considerando $e=31$ conjuntos difusos de forma triangular o L para cada uno de los tres inputs de la relación difusa de esta sección.

Tabla 12. Características de las variables raíz de la sección *d-separada*

Variable	Descripción	Unidad	Evolución temporal	Distribución estadística ^(a)	Fuente
DW_t	Diferencial de salarios	Ratio	Decreciente	U [3.70, 4.90]	Banco Mundial (2011b)
SB_t	Beneficios sociales	Dólares per cápita	Creciente	U [1850, 3600]	OCDE (2011)
MC_t	Costes de la emigración	Dólares per cápita	Decreciente	U [1500, 5000]	Jokisch y Pribilsky (2002)
PO_t	Población en origen	Habitantes (millones)	Creciente	U [7.25, 8.8]	INEC (2011b)

^(a) Distribución estadística: triangular (T), uniforme (U).

Fuente: Elaboración propia.

Tabla 13. Relaciones de dependencia difusa: características de las variables inputs

Variable	Descripción	Unidad	RD	UD	Influencia
DW_t	Diferencial de salarios	Ratio	13	[3.70, 4.90]	Positiva
SB_t	Beneficios sociales	Dólares per cápita	13	[1850, 3600]	Positiva
MC_t	Costes de la emigración	Dólares per cápita	13	[1500, 5000]	Negativa

^(a) Distribución estadística triangular (T)

Fuente: Elaboración propia.

Tabla 14. Relaciones de dependencia difusas: características de las variables outputs

Variable	Descripción	Unidad	RD	UD	Tipo	Intensidad	Prob.	Fuente
MP_t	Propensión a emigrar	Porcentaje	13	[0.01, 0.05]	Neutra	Neutra	0,98	INEC (2011b, 2010)

Fuente: Elaboración propia.

4.2.2.3. Parámetros del AEMO-RB

El ajuste de los parámetros es un punto crítico en el diseño del AEMO-RB, puesto que determina el desempeño del algoritmo para encontrar soluciones óptimas al problema de los pesos o relevancia de las variables en las relaciones difusas definidas. A pesar de que la definición de los parámetros es uno de los temas que más atención ha recibido en el campo de los algoritmos evolutivos, no existen métodos totalmente satisfactorios para definir cada uno de los parámetros iniciales del algoritmo (Lobo *et al.*, 2007).

En el caso del AEMO-RB, la determinación de los parámetros del algoritmo se ha llevado a cabo de forma experimental. Se realizaron cinco experimentos a partir de distintas semillas, diseñadas de forma aleatoria. Cada una de estas semillas (población inicial) está constituida por 100 individuos o soluciones y, a su vez, cada una de estas soluciones está compuesta de tres pesos, correspondientes a las variables de la relación difusa que define la propensión a emigrar MP_t .

Para el diseño del AEMO-RB se ha establecido como criterio de factibilidad de las soluciones que la suma de estos pesos debe ser igual a la unidad. (Ec. 30). Las soluciones no factibles son mutadas por el AEMO-RB hasta obtener una solución factible. El AEMO-RB ha sido diseñado de tal forma que el último de los pesos que compone una solución se calcula como diferencia entre la unidad y la suma de los otros componentes de la solución. Por tanto, el operador de mutación se aplica, de forma iterativa, sobre los otros componentes de la solución hasta que su suma sea

menor o igual a la unidad.

Para el criterio de parada, basado en el cálculo del ECM de la función de aptitud de cada solución w_j , se establece que cuando dicho error, denotado como $E_{FF_{jg}}$, quede $e_{FF}=1\%$ por debajo de su valor medio durante al menos 10 generaciones consecutivas, el algoritmo se detenga, obteniendo así las soluciones finales. Por otra parte, el número máximo de generaciones g_{max} se establece teniendo en cuenta la estructura de la red bayesiana o, en su caso, de la sección *d-separada* de la red para la que se ejecuta el AEMO-RB.

En cuanto a los parámetros relativos a los operadores genéticos, el AEMO-RB salvaguarda las mejores soluciones para que no sean mutadas, de tal forma que el operador elitismo sólo selecciona las dos mejores soluciones w_j^* de cada generación g , a partir de la vigésima generación. El operador genético de selección se ejecuta mediante el método de torneo binario con reemplazamiento con una probabilidad de $p_T=0,01$, mientras que para el operador de mutación la probabilidad se ha elevado hasta $p_M = 0,1$.

El AEMO-RB ha sido diseñado para que los pesos de los objetivos establecidos sobre el ECM y el error de Nakayama, que conforman la función de aptitud FF_{jg} , sean determinados de forma aleatoria por el MSCM. Sin embargo, en este diseño experimental, los pesos han sido fijados de forma predeterminada, reduciendo así la aleatoriedad del proceso y permitiendo realizar un mejor seguimiento de los resultados obtenidos. Concretamente, en los experimentos se han empleado dos combinaciones de pesos. En el primer caso, se ha asignado la misma importancia a ambos objetivos, de tal forma que $p_1=0,5$ y $p_2=0,5$. En el segundo caso, se ha optado por dar mayor relevancia al primer objetivo, otorgándole un peso de $p_1= 0,75$ frente a $p_2=0,25$, priorizando la obtención de resultados que se ajusten a los datos reales, frente a la variabilidad del proceso de simulación.

4.2.3. Estimación de pesos en las relaciones causales difusas mediante el AEMO-RB

En la Tabla 15 se muestra las combinaciones de pesos empleadas en cada uno de los experimentos realizados con el AEMO-RB, así como el número de generaciones g en las que el algoritmo converge, obteniendo como resultado el número de migrantes ecuatorianos residentes en España a finales del año 2000.

Tabla 15. Experimentos con el AEMO-RB: pesos y número de generaciones

Experimento	Soluciones	p_1	p_2	# g
1	100	0,5	0,5	52
2	100	0,5	0,5	51
3	100	0,5	0,5	59
4	100	0,75	0,25	51
5	100	0,75	0,25	13

Fuente: Elaboración propia.

Como puede observarse, el experimento número 5, a pesar de haber sido diseñado con los mismos parámetros que el experimento 4, se detiene antes, por tanto el número de generaciones en este experimento es bastante menor que en el resto.

Los resultados ofrecidos por los distintos experimentos del AEMO-RB para el primer año del periodo (2000), revelan un grado de relevancia de las variables determinantes de la propensión a remitir distinto del señalado en base al conocimiento en el Estudio 1, como puede observarse en la Tabla 16.

Tabla 16. Descriptivos para los pesos (w_i) de las variables input (DW_t , SB_t y MC_t) en la relación difusa obtenidos por AEMO-RB^(a)

Experimento	w_{DWt}			w_{SBt}			w_{MCt}		
	Media	Máximo	Mínimo	Media	Máximo	Mínimo	Media	Máximo	Mínimo
1	0,2357 (0,0276)	0,3411 (0,2304)	0,2008 (0,1197)	0,4130 (0,0245)	0,4653 (0,2409)	0,3496 (0,2195)	0,3512 (0,0241)	0,3912 (0,2021)	0,2814 (0,2326)
2	0,2739 (0,0211)	0,3475 (0,2320)	0,2374 (0,1365)	0,3688 (0,0369)	0,4381 (0,2132)	0,2733 (0,2092)	0,3573 (0,0290)	0,4222 (0,1777)	0,3049 (0,1868)
3	0,2450 (0,0289)	0,3390 (0,2456)	0,1903 (0,1394)	0,3631 (0,0175)	0,3997 (0,2234)	0,3266 (0,2185)	0,3919 (0,0240)	0,4360 (0,1935)	0,3257 (0,2435)
4	0,2756 (0,0186)	0,3209 (0,2111)	0,2328 (0,1525)	0,3784 (0,0187)	0,4147 (0,1885)	0,3358 (0,2139)	0,3460 (0,0139)	0,3808 (0,1825)	0,3208 (0,1953)
5	0,3024 (0,0193)	0,3385 (0,3385)	0,2694 (0,2694)	0,3151 (0,0230)	0,3738 (0,2399)	0,2883 (0,2348)	0,3825 (0,0180)	0,4098 (0,2516)	0,3413 (0,2331)

^(a) Las cifras entre paréntesis corresponden a la desviación estándar de los valores indicados.

Fuente: Elaboración propia.

La Tabla 17 muestra el valor medio para el stock de migrantes ecuatorianos residentes en España en $t=2000$, obtenido en cada uno de los experimentos realizados.

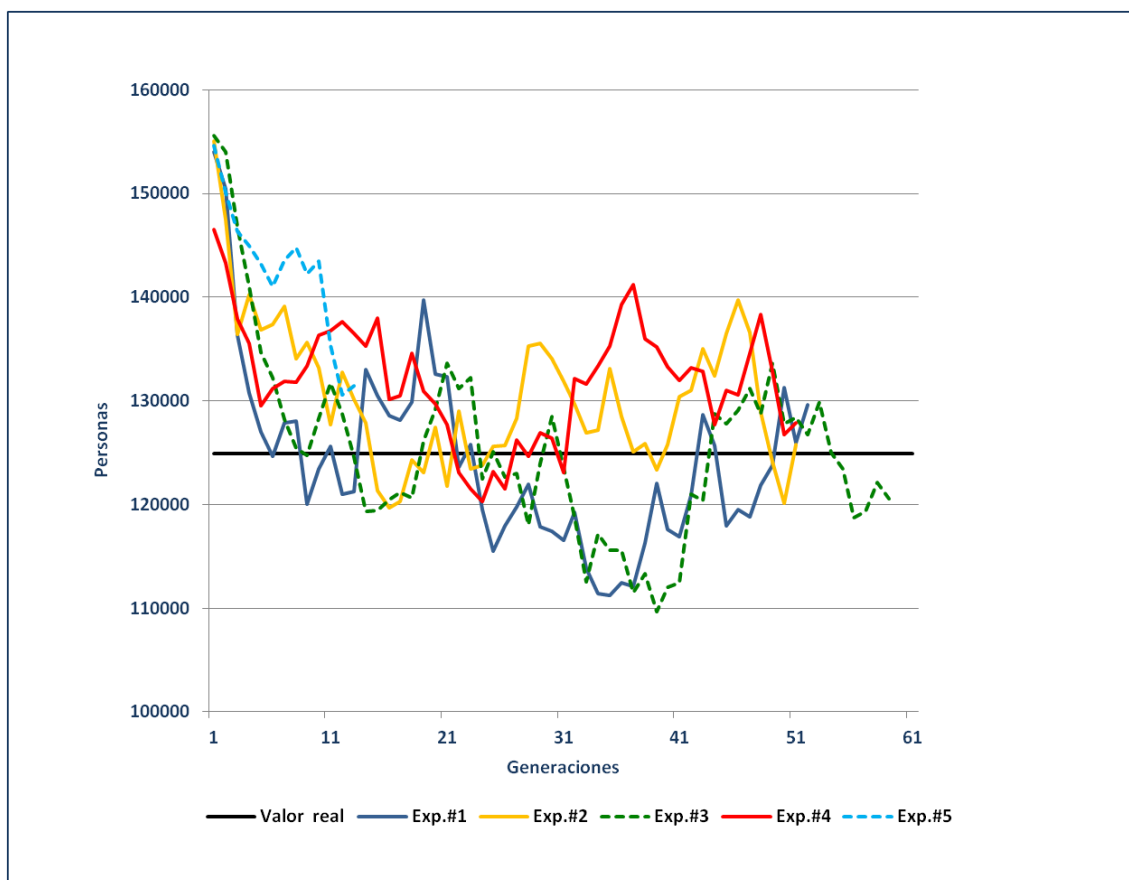
Tabla 17. Descriptivos para Nm_t (año 2000)

Experimento	Media	Desviación estándar	% ECM	% Error de Nakayama
1	124.196,55	8674,09	5,26	1,43
2	130.407,40	7013,72	4,50	1,09
3	125.391,62	9067,61	7,24	1,90
4	131.902,40	5672,69	6,88	1,22
5	142.421,28	6771,37	13,67	2,99

Fuente: Elaboración propia.

El ECM y el error de Nakayama, ambos expresados como porcentaje sobre el valor medio estimado permiten evaluar los resultados obtenidos y revelan la estabilidad del proceso de simulación. Además, en la Figura 16 se muestra como el algoritmo converge hacia el valor real del stock de población a medida que se incrementa el número de generaciones.

Figura 16. Stock de población migrante ecuatoriana en España (Año 2000): valores estimados y valores reales



Fuente: Elaboración propia.

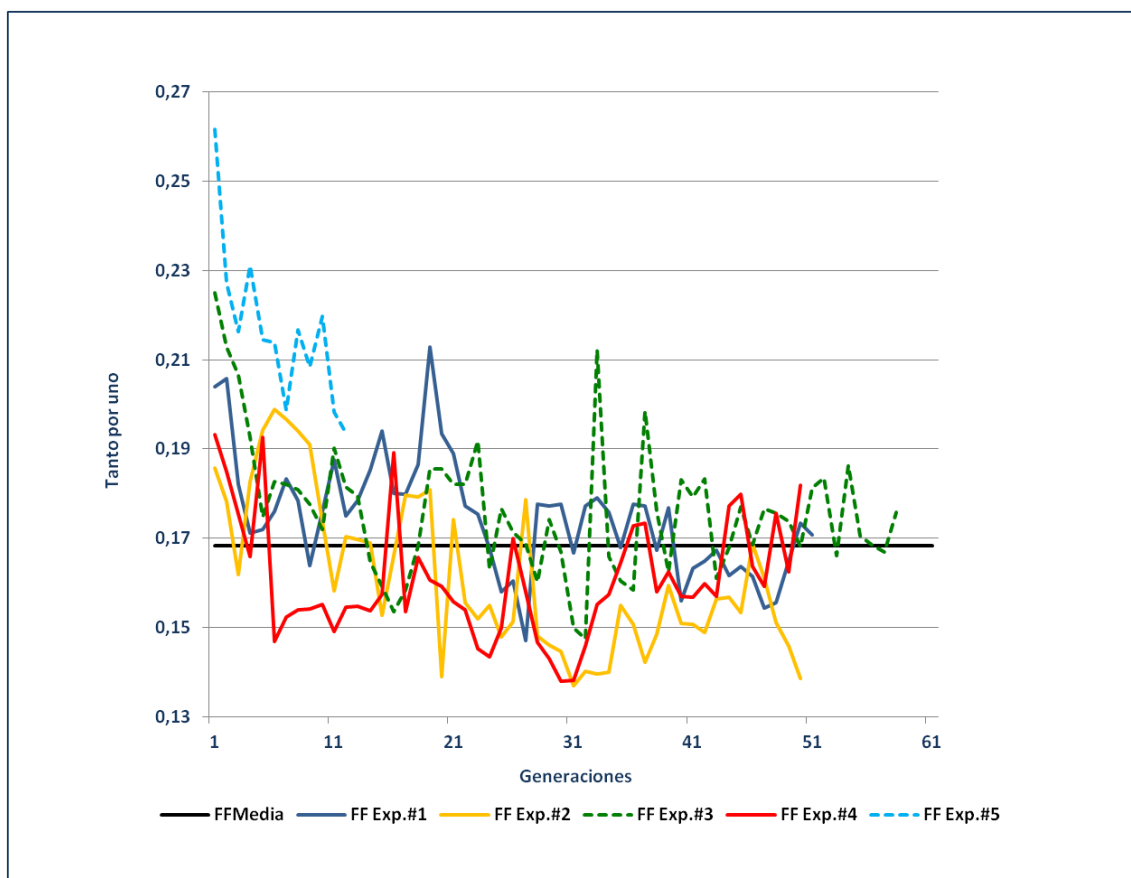
Por último, en la Tabla 18 se muestran algunos descriptivos relativos a la función de aptitud FF y para cada objetivo por separado. Como puede observarse, los valores estandarizados son bajos por lo que el MOEA converge para un número relativamente pequeño de generaciones, como es lógico. La Figura 17 muestra como los valores de la función de aptitud son cada vez menores y más cercanos a la media a medida que el número de generaciones aumenta.

Tabla 18. Descriptivos para las funciones estandarizadas [0,1, 0,9] de las soluciones eficientes

Experimento	\tilde{f}_{jg}^1		\tilde{f}_{jg}^2		FF_{jg}	
	Media	Desviación estándar	Media	Desviación estándar	Media	Desviación estándar
1	0,1483	0,1223	0,1483	0,1223	0,1748	0,1275
2	0,1460	0,1260	0,1460	0,1260	0,1614	0,1269
3	0,1476	0,1218	0,1476	0,1218	0,1761	0,1285
4	0,1446	0,1239	0,1446	0,1239	0,1607	0,1270
5	0,1798	0,1705	0,1798	0,1705	0,2188	0,1608

Fuente: Elaboración propia.

Figura 17. Evolución de la función de aptitud FF para los distintos experimentos



Fuente: Elaboración propia

V. DISCUSIÓN

5.1. El procedimiento de cálculo de la teoría T sobre la incertidumbre

En este trabajo se ha desarrollado un procedimiento de cálculo que permite manipular de forma efectiva la incertidumbre inherente a la teoría T modelizada con una red bayesiana, la cual ha sido diseñada utilizando el conocimiento experto recogido en la literatura. De esta forma, resulta posible aproximar la función de distribución de probabilidad conjunta P de dicha red tal como se planteó en el primer objetivo general.

El procedimiento diseñado ofrece una alternativa para abordar el tratamiento de variables aleatorias continuas en redes bayesianas mediante el uso de un MSMC, logrando así salvar uno de los potenciales escollos en la modelización de redes bayesianas como es la necesidad de trabajar con variables discretas. El punto de partida de este procedimiento es la identificación de la distribución estadística que determina el comportamiento de cada variable a lo largo del periodo de estudio. La selección del tipo de distribución es un aspecto crítico en este tipo de modelado (Khul *et al.*, 2008; Raychaudhuri, 2008), si bien la falta o escasez de datos ha obligado a la adopción de los supuestos pertinentes, como la asignación de distribuciones estadísticas de tipo uniforme a las variables raíz de la red bayesiana. Este inconveniente se ha tratado de reducir mediante la modelización de la evolución del comportamiento temporal de estas variables, normalmente de carácter no lineal, bien intentando ajustar los valores estimados a los datos reales disponibles, bien determinando la evolución de las variables tomando como base la información recogida en la literatura científica, tal y como ocurre en el Estudio 1 con variables como el desempleo (UD_t) o los costes de la emigración (MC_t) respectivamente.

La inclusión del MID-RB en el MSCM permite afrontar otra de las fuentes de incertidumbre de la teoría T como es la imprecisión en la definición de las relaciones causales de la red bayesiana. La hibridación de estas dos técnicas, junto con la medición de la incertidumbre, hace plenamente operacional la teoría T (Klir, 2006) que

representa la evaluación de los flujos de remesas. Su utilización convierte en funcionales aquellas relaciones causales entre variables descritas en la literatura científica, y simplifica de forma considerable el proceso de construcción de la base de conocimiento sobre la que opera un MID convencional, reduciendo el tiempo de procesamiento necesario para su ejecución.

El ajuste de los parámetros del MID-RB resulta ser un aspecto clave para el desempeño del procedimiento de cálculo propuesto, puesto que influyen en la evaluación de las relaciones causales y, por tanto, en los resultados obtenidos. La determinación de los parámetros relativos al *tipo* y la *intensidad* de la relación difusa resulta menos compleja, ya que es posible su definición a partir del conocimiento existente y, en el peor de los casos -cuando este conocimiento no está disponible, es posible asignar un valor neutro a dichos parámetros ($Q=D=0$), de tal forma que no se introduce ninguna distorsión en la evaluación de las relaciones causales correspondientes.

Sin embargo, la asignación de los *pesos*, o niveles de relevancia de las variables inputs en las relaciones difusas, representa uno de los aspectos más controvertidos en la calibración del procedimiento de cálculo, como se deduce de la comparación de los pesos estimados en los dos Estudios realizados. La determinación de los *pesos*, en base al conocimiento, presenta limitaciones en lo concerniente tanto a la definición de la distribución estadística que rige el valor de estos *pesos*, como al inconveniente que supone la determinación de los mismos a partir de un conocimiento fragmentado, de forma tal que ambas limitaciones se potencian.

El problema de la cuantificación de estos *pesos*, entendidos como niveles de relevancia, radica más en la determinación del intervalo de posibles valores que en la determinación del tipo de distribución estadística. Establecer los intervalos que estiman el *peso* relativo de una variable es una tarea difícil cuando dicha relevancia se expresa en términos semánticos o cuando, a pesar de expresarse en términos

numéricos, el conocimiento existente no contempla de forma simultánea todas las variables que han podido ser identificadas como inputs de una misma relación difusa. Este último escollo se hace más patente cuando se integra variables pertenecientes a distintos dominios de estudio, como en el ejemplo elegido en el Estudio 2, de tal forma que la asignación de *pesos* debe realizarse por secciones d-separadas de la red bayesiana diseñada que, normalmente, suelen corresponder a áreas de conocimiento concretas (Chen y Pollino, 2012). Este aspecto impide tener una visión global de la relevancia de cada variable en el conjunto del modelo, a menos que se utilicen técnicas como las presentadas en este trabajo.

El ajuste de los *pesos* de las variables de la red bayesiana, al igual que el del resto de parámetros del procedimiento de cálculo, o incluso de la propia estructura de la red o de las variables aleatorias que la conforman, puede ser mejorado mediante la iteración de sucesivas intervenciones sobre la red diseñada. Estas intervenciones deben ser, en cualquier caso, coherentes con el marco teórico existente para el problema que se pretende abordar mediante la teoría *T* propuesta. En el caso de los *pesos* de las variables, estas intervenciones han sido automatizadas con éxito mediante el diseño e implementación del AEMO-RB, tal y como se planteó en el segundo de los objetivos generales.

La estrategia de implementación del AEMO permite aproximar el conjunto de soluciones eficientes en problemas complejos, como el que nos ocupa, donde el número de soluciones factibles es infinito. El diseño del AEMO-RB presenta dos aspectos clave, por una parte el planteamiento del problema multi-objetivo y, por otra, el ajuste de los parámetros del algoritmo.

En relación al primer aspecto, el problema multi-objetivo ha sido planteado mediante la combinación de dos objetivos relativos a la estabilidad del proceso de simulación (Error de Nakayama) y a la validación de los resultados obtenidos (ECM) con la red diseñada. En este planteamiento se ha excluido como función objetivo el

índice de Shannon, a pesar de ser la medida de la incertidumbre para la teoría T desarrollada. Su inclusión como objetivo requiere de una mayor experimentación relacionada con el procedimiento de cálculo (discretización).

En relación al segundo aspecto, el ajuste de los parámetros del AEMO-RB no parece ser la causa de las discrepancias entre los pesos establecidos en base al conocimiento disponible y los estimados por el algoritmo, ya que, modificando la semilla y los pesos de los objetivos en la función de aptitud, los niveles de relevancia estimados para las variables mantienen, en general, un orden de importancia similar. Dicho orden no es totalmente contradictorio con el establecido a partir de la literatura empírica, puesto que determina una mayor relevancia de los costes de la emigración respecto al diferencial de salarios, si bien considera como tercera variable implicada en esta relación, los beneficios sociales (SB_t), con mayor relevancia que el resto de variables.

Cabe señalar que el AEMO-RB propuesto en esta investigación se encuentra en fase de desarrollo. Su diseño podría ser mejorado mediante una mayor experimentación sobre los pesos de las funciones objetivo o incluso mediante la consideración de otras propuestas para el diseño de la función de aptitud del algoritmo. A pesar de ello, los resultados obtenidos muestran que dicho procedimiento abre una nueva vía de gran potencial para la identificación de las distribuciones estadísticas que determinan los pesos de cada variable en la red bayesiana y, por tanto, permite ampliar el horizonte de la discusión teórica sobre el problema concreto al que resulte de aplicación la teoría T propuesta.

5.2. La operatividad de la teoría T : los estudios sobre flujos de remesas y población inmigrante en el corredor España-Ecuador

La propuesta metodológica de esta investigación constituye en su conjunto una teoría T sobre la incertidumbre, cuya operatividad ha sido demostrada mediante la

aplicación al fenómeno concreto de los flujos de remesas y del stock de migrantes en el corredor España-Ecuador durante el periodo 2000-2010, como se contemplaba en el tercer objetivo general.

La revisión de la literatura ha permitido desarrollar un modelo conceptual causal a partir del conocimiento sobre los flujos de remesas y sus determinantes, tanto a nivel microeconómico como macroeconómico. La fragmentación del vasto cuerpo teórico sobre éste fenómeno, según las distintas áreas de conocimiento desde las que se aborda fenómenos claramente conexos como las migraciones laborales y los flujos de remesas, ha permitido la construcción de una red bayesiana para la estimación de flujos de remesas.

La red bayesiana propuesta ha sido evaluada y validada con éxito en dos estudios, mediante la aplicación de la propuesta metodológica desarrollada en esta investigación. En el Estudio 1, la hibridación de un MSMC y el MID-RB diseñado ha permitido la aproximación de los flujos de remesas enviados desde España a Ecuador durante el periodo 2000-2010. A partir de estos resultados es posible determinar la probabilidad conjunta P de la red y medir la incertidumbre asociada al modelo, tal y como se había establecido en los objetivos de esta investigación.

En relación a los resultados estimados, un análisis detallado de los flujos medios de remesas pone de relieve que, para un primer subperiodo (2000-2005), los flujos estimados ajustan los valores registrados durante los primeros años (2000-2002), ya que muestran un bajo ECM. Por el contrario, el resto de años (2003-2005) exhibe las mayores diferencias entre flujos estimados y reales, como refleja el incremento del ECM. Si bien, cabe señalar que las diferencias observadas durante este último tramo pueden, en parte, estar motivadas por el método empleado para regionalizar los flujos de remesas anteriores a 2005. Así pues, en los datos sobre remesas proporcionados

por el Banco Central de Ecuador, se advierte de este hecho⁸. Esta sospecha se fundamenta en la existencia de algunos datos anómalos, por ejemplo, durante los primeros años (2000-2002) la remesa anual per cápita descende se sitúa en torno a los 1.500 dólares, y a partir de 2005 (año en el que empieza a llevarse un control del origen de las remesas) esta cantidad se duplica hasta alcanzar los 3.200 dólares. Este salto también resulta llamativo, pues entre 2004 y 2005 la población ecuatoriana empadronada en España apenas registra una variación de 20.000 individuos en una población de cerca de 400.000.

En cuanto al segundo subperiodo (2006-2010), como ya se ha mencionado, los flujos estimados replican el máximo alcanzado por los flujos registrados, cercano a los 1288 MM US\$, de tal forma que el ECM comienza a descender. Sin embargo, en el los últimos años (2008-2010), los flujos estimados son inferiores a los registrados, ya que el modelo detecta una reducción más acusada que la producida en los flujos de remesas registrados, incrementándose el ECM correspondiente a partir del año 2009. En este caso las diferencias pueden ser atribuibles a varias causas, entre las que cabe señalar, un cambio en los pesos asignados a las variables del problema, o la aparición de alguna nueva variable explicativa, dada la magnitud de la crisis económica-financiera durante dicho periodo, que no haya sido recogido en el modelo.

El control de variabilidad, establecido sobre el MSMC, ha puesto de manifiesto la estabilidad del proceso de simulación, puesto que los valores mostrados por el error de Nakayama fueron muy bajos. Este hecho revela que las diferencias entre los valores estimados y los valores registrados de remesas no se deben a un mal desempeño del procedimiento de cálculo diseñado, sino que resulta necesario un mejor ajuste de los parámetros del modelo. Este ajuste podría ser mejor a medida que aumenta los datos

⁸“El Banco Central del Ecuador, no cuenta con un desglose por país de procedencia desde 1993 al 2004. Sin embargo con el fin de establecer la trayectoria de las remesas en estos años se realizó una aproximación con base a información publicada por varias revistas especializadas, estudios realizados y entrevistas con varios gerentes de las empresas Courier que iniciaron el pago de remesas”(Banco Central de Ecuador, 2012).

disponibles, permitiendo afinar el tipo de distribución estadística que modela el comportamiento de las variables raíces. También podría ser mejorado mediante sucesivas intervenciones sobre el modelo diseñado. Esta última alternativa demanda una gran cantidad de recursos de computación y tiempo, dada la complejidad del modelo y la gran cantidad de combinaciones posibles, siendo parcialmente abordada mediante el diseño del AEMO-RB.

La reducción de la incertidumbre asociada a la probabilidad conjunta de la red bayesiana ha sido evaluada mediante la Entropía de Shannon. El método de cálculo seleccionado para este índice exige de la discretización de los resultados en un número de intervalos no prefijado que, como muestran los resultados, influye de manera decisiva en el valor de este índice. Por ello, la teoría T propuesta puede ser mejorada mediante el uso de una medida de la incertidumbre para variables aleatorias continuas, como la medida de entropía relativa propuesta por Kullback y Leibler (1951). Esta mejora debe ser valorada puesto que su aplicación exige la asunción a priori del tipo de función de distribución de probabilidad para la variable o variables sobre las que se desee medir la entropía.

En el Estudio 2, el AEMO-RB diseñado para la estimación de los *pesos* de las variables ha sido aplicado a una sección *d-separada* de la red bayesiana, relativa al stock de población migrante ecuatoriana en España. La selección de esta sección se debe a los motivos ya señalados: disponibilidad de datos sobre el stock de inmigrantes y comparación con los pesos establecidos en el Estudio 1.

Como ya se ha señalado y se esperaba, los pesos estimados por el AEMO-RB difieren de los establecidos en el Estudio 1 en base al conocimiento. Esta diferencia puede ser explicada en base a la información empleada en el Estudio 1 para la calibración de los *pesos* de la red. Dado que no existe ningún estudio empírico que considerara, de forma conjunta, las tres variables implicadas en la determinación de la propensión a emigrar a España de la población ecuatoriana, el nivel de relevancia de la

variable relativa de los beneficios sociales (SB_t), se estableció atendiendo a los resultados de estudios empíricos en otras zonas geográficas, mientras que para la determinación de los pesos de las otras dos variables, diferencial de salarios (DW_t) y costes de la migración (MC_t), se disponía de la referencia establecida en (Bertoli *et al.*, 2011).

Los resultados obtenidos por el AEMO-RB asignan, en general, una mayor importancia relativa a los beneficios sociales como variable determinante de la propensión a emigrar en el corredor estudiado, confirmando la hipótesis de atracción del estado del bienestar (WMH, por sus siglas en inglés). Estos resultados difieren de aquellos estudios empíricos en los que España aparece como destino migratorio. El estudio de Giulietti *et al.* (2013) soporta la WMH para inmigrantes no pertenecientes a la Unión Europea durante el periodo 1993-2008. Sin embargo, en este estudio se emplea, como variable proxy para esta hipótesis, los subsidios por desempleo que fueron descartados en nuestro análisis, dada la dificultad de la población inmigrante para generar este tipo de derechos. Por su parte, el trabajo de Pedersen *et al.* (2008), que analiza los factores determinantes de la inmigración para los países de la OCDE durante el periodo 1990-2000, rechaza la WMH y pone de relieve la importancia de factores económicos como el diferencial de salarios.

Los resultados obtenidos por el AEMO-RB, el uso de distintas semillas y la estabilidad del proceso de simulación, apuntan la posibilidad de que las diferencias en el nivel de relevancia de las variables pueden estar ocasionadas por la caracterización de las variables aleatorias, por la ausencia de alguna otra variable explicativa de la propensión a emigrar o por aspectos socioeconómicos del ámbito geográfico estudiado, ya que las teorías económicas que soportan el marco conceptual son de carácter generalista.

5.3. Difusión de la propuesta metodológica

Esta investigación ha sido parcialmente desarrollada en el marco del proyecto de investigación TIN2011-22794 del Ministerio Español, Comisión de Ciencia y Tecnología (MICYT), fondos FEDER y el proyecto P2011-TIC-7508 de la Junta de Andalucía (España). La producción científica derivada de la misma se ha difundido en 5 congresos:

- Campoy-Muñoz, P.; García-Alonso, C.; Salazar-Ordoñez, M. (Junio 2013): Dealing with migratory flows under uncertainty: a multi-objective optimization approach. 22nd International Conference on Multiple Criteria Decision Making (Málaga, España).
- Campoy-Muñoz, P.; Gutiérrez-Peña, P.A.; Hervás-Martínez, C. (Junio 2013): Addressing remitting behavior using an ordinal classification approach. 5th. International Work-Conference on the interplay between natural and artificial computation (Mallorca, España).
- Campoy-Muñoz, P.; Salazar-Ordoñez, M.; García-Alonso (Mayo 2012): Remittance flows estimation hybridizing a multilevel fuzzy system and a simulation model. 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD'12) (Chongqing, China).
- Campoy-Muñoz, P.; Salazar-Ordoñez, M.; García-Alonso (Octubre de 2011) Comunicación: A bayesian model for remittances and gender: the Spain-Ecuador corridor. International Conference “New debates on gender, migration and development: linking production and reproduction chains in contemporary mobility” (Menorca, España).
- Campoy-Muñoz, P.; Salazar-Ordoñez, M.; García-Alonso (Junio de 2011): Aproximación a un modelo causal para la estimación de remesas: factores determinantes y sus relaciones. III International Conference about Migrations. (Oporto, Portugal).

Y ha sido publicada en:

- García-Alonso, C.; Campoy-Muñoz, P.; Salazar-Ordoñez, M. A Multi-Objective Evolutionary Algorithm for enhancing Bayesian Networks hybrid-based modeling. Computers and Mathematics with Applications. DOI: <http://dx.doi.org/10.1016/j.camwa.2013.01.029> (Anexo X)
- Campoy-Muñoz, P.; Salazar-Ordoñez, M.; García-Alonso, C. A bayesian network for remittances and gender in the Spain-Ecuador corridor. The International Handbook on Gender, Migration and Transnationalism: Global and Development Perspectives. Edwar Elgar Publisher (En imprenta).
- Campoy-Muñoz, P.; Salazar-Ordoñez, M.; García-Alonso, C. A hybrid-based model for estimating remittance flows: the Spain-Ecuador corridor. The European Journal of Development Research (En revisión).

Los conocimientos y habilidades desarrollados a lo largo de esta investigación han sido de utilidad para la el desarrollo de diversos trabajos publicados en revistas científicas:

- Fernandez-Navarro F.; Campoy-Muñoz, P.; De la Paz-Marín, M.; Hervás-Martínez, C; Yao, X. Addressing the EU sovereign ratings using an ordinal regression approach. IEEE Systems, Man, and Cybernetics, Part C: applications and reviews. DOI: <http://dx.doi.org/10.1109/TSMCC.2013.2247595>.
- De la Paz-Marín, M.; Campoy-Muñoz, P.; Hervás-Martínez, C. Non-linear multiclassifier model based on Artificial Intelligence to predict Research and Development Performance in European Countries. Technological Forecasting & Social Change, 79 (9), pp 1731-45.

VI. CONCLUSIONES Y LÍNEAS FUTURAS DE INVESTIGACION

6.1. Conclusiones

El procedimiento de cálculo, resultante de la hibridación de técnicas de simulación e inteligencia artificial, ha sido aplicado con éxito para la evaluación de redes bayesianas diseñadas en base a conocimiento experto, como se planteaba en las dos primeras hipótesis de esta investigación.

El MSMC, que soporta el procedimiento de cálculo diseñado en esta investigación, constituye una nueva alternativa para la evaluación de redes bayesianas en las que las variables aleatorias del dominio de estudio presentan una naturaleza continua, de forma que, no se introduce ningún potencial sesgo en el procedimiento de evaluación del modelo como consecuencia de la discretización de las variables que lo conforman. Además, la posibilidad de modelar la evolución temporal del comportamiento de las variables ofrece una primera aproximación al modelado de fenómenos dinámicos.

La hibridación del MSMC y del MID-RB permite la evaluación de redes bayesianas en las que las relaciones causales son formalizadas atendiendo al tipo de incertidumbre identificada en el conocimiento disponible para su diseño. De esta forma, se amplía el campo de modelado de las redes bayesianas ya que, junto a la relaciones convencionales de tipo probabilístico, el procedimiento de cálculo hace funcionales relaciones de carácter algebraico y, especialmente, aquellas de naturaleza imprecisa. La formalización de este tipo de relaciones ofrece una vía de gran potencial y robustez para el modelado, mediante redes bayesianas, de aquellos fenómenos en los que la ausencia o escasez de los datos impide el uso de métodos probabilísticos convencionales.

La incorporación del AEMO-RB en el procedimiento de cálculo introduce una importante novedad en la evaluación de las redes bayesianas. Además de reducir parte del potencial grado de subjetividad, generado por el uso del conocimiento en el diseño

y evaluación de la red, el AEMO-RB permite ampliar la discusión teórica con base empírica en áreas de conocimiento donde la escasez de datos puede resultar una seria limitación para estudios aplicados, ya que posibilita aproximar la distribución estadística que caracteriza el nivel de relevancia de las variables de la red.

Por otra parte, la teoría T propuesta en esta investigación resulta ser plenamente operacional en aquellos problemas de decisión caracterizados por una elevada fragmentación del conocimiento existente y/o por la escasez de datos. Así pues, los resultados de la aplicación empírica realizada con los flujos de remesas constatan que es posible cuantificar la incertidumbre presente en este tipo de problemas, tal y como se proponía en la tercera hipótesis.

La cuantificación de la incertidumbre sobre el fenómeno objeto de estudio y de su reducción, mediante el procedimiento de ajuste de la función P a través de sucesivas iteraciones del procedimiento de cálculo tras realizar intervenciones en la red, se ha llevado a cabo mediante la Entropía Shannon. Sin embargo, se ha detectado que el método seleccionado para calcular este índice no es totalmente satisfactorio, puesto que el número de intervalos utilizados en dicho cálculo influye de forma decisiva en su valor. Esta limitación, cuya mejora supone una línea futura de ampliación del trabajo, no resta valor a la propuesta para la medición de la información generada (entendida como reducción de la incertidumbre) mediante intervenciones en el modelo, determinado así su validez como herramienta de apoyo al problema de decisión que se pretenda modelar.

La utilización de la teoría T en el estudio de los flujos de remesas ha permitido demostrar empíricamente la aplicabilidad de dicha teoría y sus posibilidades como herramienta para los sistemas de apoyo a la decisión de aquellos agentes interesados en los mismos. Como se ha podido observar, la herramienta permite aproximar los flujos de remesas recibidos en una economía, cuya determinación resulta fundamental de cara al diseño de políticas económicas que potencien los efectos positivos de estos

flujos, o atenúen e incluso neutralicen los efectos negativos, sobre las economías de los países receptores.

La utilidad de esta herramienta se extiende también a la evaluación de los efectos causados por intervenciones sobre las variables determinantes de los flujos de remesas. La aproximación del nivel de relevancia de las variables de la red bayesiana es de suma importancia en este tipo de herramientas, puesto que permite identificar vías de acción para modular los flujos de remesas recibidos en una economía. Las variables identificadas como claves para los flujos de remesas en un determinado contexto geográfico pueden ser objeto de programas, medidas o acciones públicas o privadas encaminadas a situar dicha variable en un determinado valor que, de acuerdo con las estimaciones de la red, provoque el efecto deseado sobre los mismos.

Por último, la flexibilidad del modelo causal, sobre el que se sustenta la herramienta, permite la incorporación de nuevas variables o relaciones causales que reflejen la evolución del conocimiento existente sobre el fenómeno de los flujos de remesas, las particularidades existente en una zona geográfica concreta o la exploración de nuevas relaciones o de nuevas variables sobre dichos flujos.

En definitiva, el procedimiento híbrido diseñado abre una importante vía para la modelización de redes bayesianas diseñadas en base al conocimiento en la que, además de ser posible aproximar la función de probabilidad conjunta P mediante una adecuada caracterización de los distintos tipos de incertidumbre, la determinación de los niveles de relevancia de las variables de la red aporta información relevante sobre el fenómeno objeto de estudio, permitiendo, en algunos casos, ampliar los planteamientos teóricos existentes. Además, la generación de información (útil) para la toma de decisiones, en problemas caracterizados por un conocimiento fragmentado y/o datos escasos, utilizando la teoría T desarrollada en esta investigación, convierte dicha propuesta en un prototipo de herramienta de gran utilidad y potencialidad como sistema de apoyo a la decisión.

6.2. Líneas futuras de desarrollo de la investigación

La aplicación de la teoría T propuesta a un fenómeno complejo, como es la estimación de los flujos de remesas, ha puesto de relieve puntos de mejora en el procedimiento de cálculo diseñado. En relación a éste, las líneas de investigación futura se articulan en torno a dos propuestas relativas al diseño del AEMO-RB y al uso de la red bayesiana como metodología para la modelización de la incertidumbre.

En el primer caso, la potencialidad demostrada por el AEMO-RB para la determinación de los *pesos* de las variables y , por tanto, para la automatización parcial del proceso de ajuste de la probabilidad de la red, anima a la mejora de su diseño mediante la consideración de configuraciones alternativas de la función de aptitud, así como a una mayor experimentación que permita establecer los intervalos de valores óptimos para los pesos de las funciones objetivo. También, habría que considerar la posibilidad de incluir, en el planteamiento del problema de optimización, una función objetivo relativa a la medida de incertidumbre de los resultados de la red, considerando en todo caso la naturaleza de las variables aleatorias del modelo.

En cuanto a la segunda propuesta, cabe señalar que la estructura del MSMC ha hecho posible la modelización del comportamiento del fenómeno objeto de estudio a lo largo tiempo. Este hecho, que representa una evolución respecto a la estimación individual de la red para cada uno de los años de un periodo, abre una importante línea de mejora para la modelización dinámica de fenómenos complejos. En este sentido, una futura línea de investigación consistiría en sustituir la red bayesiana convencional, empleada en la teoría T propuesta, por una red bayesiana dinámica (modelo declaratorio), que explícitamente considere relaciones causales de naturaleza temporal, derivándose la necesidad de modelar dichas relaciones y de adaptar, si fuera necesario, el procedimiento de cálculo propuesto.

REFERENCIAS BIBLIOGRÁFICAS

- Acosta, P.A., Lartey, E.K.K., Mandelman, F.S., 2009. Remittances and the dutch disease. *Journal of International Economics* 79, 102–116.
- Adams, R.H., 2009. The determinants of international remittances in developing countries. *World Development* 37, 93–103.
- Adams, R.H., Cuecuecha, A., 2010. Remittances, household expenditure and investment in Guatemala. *World Development* 38, 1626–1641.
- Adams, R.H., Page, J., 2005. Do international migration and remittances reduce poverty in developing countries? *World Development* 33, 1645–1669.
- Al-Sakka, R., ap Gwilym, O., 2010. Split sovereign ratings and rating migrations in emerging economies. *Emerging Markets Review* 11, 79–97.
- Amuedo-Dorantes, C., De La Rica, S., 2007. Labour market assimilation of recent immigrants in Spain. *British Journal of Industrial Relations* 45, 257–284.
- Amuedo-Dorantes, C., Pozo, S., 2006a. Migration, remittances, and male and female employment patterns. *American Economic Review* 96, 222–226.
- Amuedo-Dorantes, C., Pozo, S., 2006b. The time pattern of remittances: evidence from Mexican migrants. *Well-being and Social Policy* 2, 49–66.
- Amuedo-Dorantes, C., Pozo, S., 2013. Remittances and portfolio values: an inquiry using immigrants from Africa, Europe, and the Americas. *World Development* 41, 83–95.
- Aydas, O.T., Metin-Ozcan, K., Neyapti, B., 2005. Determinants of workers' remittances: the case of Turkey. *Emerging Markets Finance and Trade* 41, 53–69.
- Ball, C.P., Lopez, C., Reyes, J., 2013. Remittances, inflation and exchange rate regimes in small open economies. *The World Economy* 36, 487–507.
- Banco Central de Ecuador, 2010. La economía ecuatoriana luego de 10 años de dolarización (Documento de trabajo). Banco Central de Ecuador, Quito.
- Banco Central de Ecuador, 2011. Estadísticas sobre sector monetario: tasas de interés. URL <http://www.bce.fin.ec/frame.php?CNT=ARB0000061> (Último acceso, Marzo 2012).
- Banco Central de Ecuador, 2012. Estadística sobre remesas. URL <http://www.bce.fin.ec/frame.php?CNT=ARB0000985> (Último acceso, Marzo 2012).
- Banco de España, 2011. Estadísticas sobre tipos de interés y tipos de cambio. URL <http://www.bde.es/webbde/es/estadis/infoest/tipos/tipos.html> (Último acceso, Marzo 2012).
- Banco Mundial, 2011a. Remittances Prices Worldwide. URL <http://remittanceprices.worldbank.org/> (Último acceso, Marzo 2012).

- Banco Mundial, 2011b. Data. GNI per capita at power purchasing parity (current international US\$). URL <http://data.worldbank.org/indicator/NY.GNP.PCAP.PP.CD> (Último acceso, Marzo 2012).
- Barth, E., Bratsberg, B., Raaum, O., 2004. Identifying earnings assimilation of immigrants under changing macroeconomic conditions. *Scandinavian Journal of Economics* 106, 1–22.
- Baumgartner, K., Ferrari, S., Palermo, G., 2008. Constructing Bayesian networks for criminal profiling from limited data. *Knowledge-Based Systems* 21, 563–572.
- Bertoli, S., Fernández-Huertas Moraga, J., 2012. Visa policies, networks and the cliff at the border (Documento de trabajo No. 7094). Instituto para el Estudio del Trabajo (IZA), Bonn.
- Bertoli, S., Moraga, J.F.-H., Ortega, F., 2011. Immigration policies and the Ecuadorian exodus. *The World Bank Economic Review* 25, 57–76.
- Biller, B., Gunes, C., 2010. Introduction to simulation input modeling, in: *Proceedings of the 2010 Winter Simulation Conference (WSC'10)*. IEEE, pp. 49–58.
- Blanchflower, D.G., Oswald, A.J., 1990. The wage curve. *Scandinavian Journal of Economics* 92, 215–35.
- Bollard, A., McKenzie, D., Morten, M., Rapoport, H., 2011. Remittances and the brain drain revisited: the microdata show that more educated migrants remit more. *The World Bank Economic Review* 25, 132–156.
- Borjas, G.J., 1989. Economic theory and international migration. *International Migration Review* 23, 457–485.
- Borjas, G.J., 1999. Immigration and welfare magnets. *Journal of Labor Economics* 17, 607–637.
- Borjas, G.J., 2000. *Issues in the economics of immigration*. The University of Chicago Press, Chicago.
- Borjas, G.J., 2001. *Heaven's door: immigration policy and the american economy*. Princeton University Press, Princeton.
- Bosma, R., Kaymak, U., van den Berg, J., Udo, H., Verreth, J., 2010. Using fuzzy logic modelling to simulate farmers' decision-making on diversification and integration in the Mekong Delta, Vietnam. *Soft Computing* 15, 295–310.
- Bourdet, Y., Falck, H., 2006. Emigrants' remittances and dutch disease in Cape Verde. *International Economic Journal* 20, 267–284.
- Bromley, J., Jackson, N.A., Clymer, O.J., Giacomello, A.M., Jensen, F.V., 2005. The use of Hugin® to develop Bayesian networks as an aid to integrated water resource planning. *Environmental Modelling & Software* 20, 231–242.

- Brown, R.P., 1997. Estimating remittance functions for Pacific Island migrants. *World Development* 25, 613–626.
- Brown, R.P.C., Connell, J., Jimenez-Soto, E.V., 2013. Migrants' remittances, poverty and social protection in the South Pacific: Fiji and Tonga. *Population, Space and Place* in press.
- Buch, C.M., Kuckulenz, A., 2010. Worker remittances and capital flows to developing countries. *International Migration* 48, 89–117.
- Calero, C., Bedi, A.S., Sparrow, R., 2009. Remittances, liquidity constraints and human capital investments in Ecuador. *World Development* 37, 1143–1154.
- Campoy-Muñoz, P., Gutiérrez, P.A., Hervás-Martínez, C., 2013. Addressing remitting behavior using an ordinal classification approach, in: *Proceedings of the 5th International Work-Conference on the Interplay Between Natural and Artificial Computation (IWINAC'2013)*, *Lectures Notes in Computer Science*. Springer, pp. 326–331.
- Cano, J.E., Hernández, L.D., Moral, S., 1996. Importance sampling algorithms for the propagation of probabilities in belief networks. *International Journal of Approximate Reasoning* 15, 77–92.
- Carling, J., 2008. The determinants of migrant remittances. *Oxford Review of Economic Policy* 24, 582–599.
- Celikyilmaz, A., Türksen, I.B., 2009. Modeling uncertainty with fuzzy logic: with recent theory and applications, *Studies in Fuzziness and Soft Computing*. Springer-Verlag Berlin Heidelberg, Berlin.
- Chami, R., Barajas, A., Cosimano, T., Fullenkamp, C., Gapen, M., Montiel, P.J., 2009. Macroeconomic consequences of remittances (Documento de trabajo No. 259). Fondo Monetario Internacional.
- Chami, R., Hakura, D.S., Montiel, P.J., 2012. Do worker remittances reduce output volatility in developing countries? *Journal of Globalization and Development* 3, in press.
- Charniak, E., 1991. Bayesian networks without tears. *AI magazine* 12, 50–63.
- Chen, S.H., Pollino, C.A., 2012. Good practice in Bayesian network modelling. *Environmental Modelling & Software* 37, 134–145.
- Chiswick, B., 1999. Are immigrants favorably self-selected? *American Economic Review* 89, 181–185.
- Chiswick, B.R., 1978. The effect of americanization on the earnings of foreign-born men. *Journal of Political Economy* 86, 897–921.
- Chiswick, B.R., Cohen, Y., Zach, T., 1997. The labor market status of immigrants: effects of the unemployment rate at arrival and duration of residence. *Industrial and*

- Labor Relations Review 50, 289–303.
- Chiswick, B.R., Miller, P.W., 2009. The international transferability of immigrants' human capital. *Economics of Education Review* 28, 162–169.
- Chiu, S.L., 1997. An efficient method for extracting fuzzy classification rules from high dimensional data. *Journal of Advanced Computational Intelligence* 1, 31–36.
- Choquet, G., 1953. Theory of capacities. *Annales De L'Institut Fourier* 5, 131–295.
- Clark, X., Hatton, T.J., Williamson, J.G., 2004. What explains emigration out of Latin America? *World Development* 32, 1871–1890.
- Cobb, B.R., Rumí, R., Salmerón, A., 2005. Modeling conditional distributions of continuous variables in Bayesian networks, in: Famili, A.F., Kok, J.N., Peña, J.M., Siebes, A., Feelders, A. (Eds.), *Advances in Intelligent Data Analysis VI, Lectures Notes in Computer Science*. Springer, pp. 36–45.
- Coello, C.A.C., Lamont, G.B., 2004. *Applications of Multi-Objective Evolutionary Algorithms, Advances in Natural Computation*. World Scientific Publishing Co., Singapur.
- Coello, C.A.C., Lamont, G.B., van Veldhuizen, D.A., 2007. *Evolutionary algorithms for solving multi-objective problems, 2nd ed, Genetic and Evolutionary Computation Series*. Springer-Verlag New York Inc., Nueva York.
- Combs, W.E., Andrews, J.E., 1998. Combinatorial rule explosion eliminated by a fuzzy rule configuration. *IEEE Transactions on Fuzzy Systems* 6, 1–11.
- Cooper, G.F., 1990. The computational complexity of probabilistic inference using bayesian belief networks. *Artificial Intelligence* 42, 393–405.
- Cox, E., 2005. *Fuzzy modeling and genetic algorithms for data mining and exploration*. Morgan Kaufmann, San Francisco.
- Darwiche, A., 2001. Recursive conditioning. *Artificial Intelligence* 126, 5–41.
- Darwiche, A., 2010. Bayesian networks. *Communications of the Association for Computing Machinery (ACM)* 53, 80–90.
- Das, I., Dennis, J.E., 1997. A closer look at drawbacks of minimizing weighted sums of objectives for Pareto set generation in multicriteria optimization problems. *Structural Optimization* 14, 63–69.
- De Giorgi, G., Pellizzari, M., 2009. Welfare migration in Europe. *Labour Economics* 16, 353–363.
- De Haas, H., 2010. Migration and development: a theoretical perspective. *International Migration Review* 44, 227–264.
- Dechter, R., 1999. Bucket elimination: A unifying framework for reasoning. *Artificial Intelligence* 113, 41–85.

- Druzdel, M.J., van der Gaag, L.C., 2000. Building probabilistic networks: "Where do the numbers come from?" guest editors' introduction. *IEEE Transactions on Knowledge and Data Engineering* 12, 481–486.
- Dubois, D., Prade, H., 2010. Formal representations of uncertainty, in: Bouyssou, D., Dubois, D., Pirlot, M., Prade, H. (Eds.), *Decision-making Process: Concepts and Methods*. ISTE Ltd & John Wiley & Sons, Inc, pp. 85–156.
- Durand, J., Parrado, E.A., Massey, D.S., 1996. Migradollars and development: a reconsideration of the Mexican case. *International Migration Review* 30, 423–444.
- Dustmann, C., Mestres, J., 2010. Remittances and temporary migration. *Journal of Development Economics* 92, 62–70.
- Echazarra, A., 2010. Accounting for the time pattern of remittances in the Spanish context (Documento de trabajo No. 553). Fundación de Cajas de Ahorros (FUNCAS), Madrid.
- Elbadawi, I., Rocha, R. de R., 1992. Determinants of expatriate workers' remittances in North Africa and Europe (Documento de trabajo No. 1038). Banco Mundial, Washington.
- Epstein, G.S., 2008. Herd and network effects in migration decision-making. *Journal of Ethnic and Migration Studies* 34, 567–583.
- Faini, R., 2007. Remittances and the brain drain: do more skilled migrants remit more? *World Bank Economic Review* 21, 177–191.
- Fernández, C., Ortega, C., 2008. Labor market assimilation of immigrants in Spain: employment at the expense of bad job-matches? *Spanish Economic Review* 10, 83–107.
- FLACSO, 2006. Ecuador: la migración internacional en cifras. FLACSO Ecuador, Quito.
- FMI, 2009. International transactions in remittances guide for compilers and users. Fondo Monetario Internacional, Washington.
- Fonseca, C.M., Fleming, P.J., 1995. An overview of evolutionary algorithms in multiobjective optimization. *Evolutionary Computation* 3, 1–16.
- Frankel, J., 2011. Are bilateral remittances countercyclical? *Open Economies Review* 22, 1–16.
- Freund, C., Spatafora, N., 2008. Remittances, transaction costs, and informality. *Journal of Development Economics* 86, 356–366.
- Fung, R.M., Chang, K.-C., 1990. Weighing and integrating evidence for stochastic simulation in bayesian networks, in: *Proceedings of the 5th Annual Conference on Uncertainty in Artificial Intelligence (UAI'89)*. North-Holland Publishing Co., pp. 209–220.

- Funkhouser, E., 2006. The effect of emigration on the labor market outcomes of the sender household: a longitudinal approach using data from nicaragua. *Well-Being and Social Policy* 2, 5–25.
- García-Alonso, C.R., Arenas-Arroyo, E., Pérez-Alcalá, G.M., 2012. A macro-economic model to forecast remittances based on Monte-Carlo simulation and artificial intelligence. *Expert Systems with Applications* 39, 7929–7937.
- García-Alonso, C.R., Campoy-Muñoz, P., Salazar Ordoñez, M., in press. A Multi-Objective Evolutionary Algorithm for enhancing Bayesian Networks hybrid-based modeling. *Computers & Mathematics with Applications*.
- García-Alonso, C.R., Pérez-Alcalá, G.M., 2008. Generating artificial populations using a multi-level fuzzy inference engine, in: *Proceedings of the 2008 Winter Simulation Conference (WSC'08)*. IEEE, pp. 2801–2810.
- Gegov, A., Gopalakrishnan, N., 2008. Advanced inference in fuzzy systems by rule base compression. *Mathware & Soft Computing* 14, 201–216.
- Gibson, J., McKenzie, D., 2012. The economic consequences of “brain drain” of the best and brightest: microeconomic evidence from five countries. *The Economic Journal* 122, 339–375.
- Giiven, M.K., Passino, K.M., 2001. Avoiding exponential parameter growth in fuzzy systems. *IEEE Transactions on Fuzzy Systems* 9, 194–199.
- Gilks, W.R., Richardson, S., Spiegelhalter, D.J., 1995. *Markov chain Monte Carlo in practice*, Interdisciplinary Statistics. Chapman & Hall/CRC, Cambridge.
- Giulietti, C., Guzi, M., Kahanec, M., Zimmermann, K.F., 2013. Unemployment benefits and immigration: evidence from the EU. *International Journal of Manpower* 34, 24–38.
- Grigorian, D.A., Melkonyan, T.A., 2011. Destined to receive: the impact of remittances on household decisions in Armenia. *Review of Development Economics* 15, 139–153.
- Gupta, S., Pattillo, C.A., Wagh, S., 2009. Effect of remittances on poverty and financial development in Sub-Saharan Africa. *World Development* 37, 104–115.
- Hagen-Zanker, J., Siegel, M., 2007. *The determinants of remittances: a review of the literature* (Documento de trabajo No. 3). Maastricht Graduate School of Governance, Maastricht.
- Halgamuge, S.K., Glesner, M., 1994. Neural networks in designing fuzzy systems for real world applications. *Fuzzy Sets and Systems* 65, 1–12.
- Halmos, P.R., 1950. *Measure Theory*. Princeton, Nueva Jersey.
- Harris, J.R., Todaro, M.P., 1970. Migration, unemployment and development: a two-sector analysis. *The American Economic Review* 60, 126–142.

- Higgins, M.L., Hysenbegasi, A., Pozo, S., 2004. Exchange-rate uncertainty and workers' remittances. *Applied Financial Economics* 14, 403–411.
- Holst, E., Schäfer, A., Schrooten, M., 2012. Gender and remittances: evidence from Germany. *Feminist Economics* 18, 201–229.
- Hu, X., Hui, W., Shuo, W., 2007. Using expert's knowledge to build bayesian networks, in: *Proceedings of the International Conference on Computational Intelligence and Security Workshops (CISW'07)*. IEEE, pp. 220–223.
- Hunter, P.R., Pond, K., Jagals, P., Cameron, J., 2009. An assessment of the costs and benefits of interventions aimed at improving rural community water supplies in developed countries. *Science of The Total Environment* 407, 3681–3685.
- Hysenbegasi, A., Pozo, S., 2006. Workers' remittances and currency crises. *Well-being and Social Policy Journal/Bienestar y Política Social* 2, 67–88.
- INE, 2008. Encuesta Nacional de Inmigrantes 2007. URL <http://www.ine.es/jaxi/menu.do?type=pcaxis&path=%2Ft20%2Fp319&file=inebase&L=0> (Último acceso, Marzo 2012).
- INE, 2009. Encuesta Nacional de Inmigrantes 2007: una monografía. Instituto Nacional de Estadística, Madrid.
- INE, 2010. Padrón municipal. URL <http://www.ine.es/jaxi/menu.do?type=pcaxis&path=/t20/e245/&file=inebase> (Último acceso, Marzo 2012).
- INE, 2011a. Encuesta de Población Activa. URL http://www.ine.es/jaxi/menu.do?type=pcaxis&path=/t22/e308_mnu&file=inebase&L=0 (Último acceso, Marzo 2012).
- INE, 2011b. Encuesta de estructura salarial. URL <http://www.ine.es/jaxi/menu.do?type=pcaxis&path=/t22/p133&file=inebase&L=0> (Último acceso, Marzo 2012).
- INEC, 2008. Caracterización socio-demográfica y económica de las y los emigrantes ecuatorianos (Documento de trabajo). Instituto Nacional de Estadística y Censos (INEC), Quito.
- INEC, 2010. Ecuador-Estadística de Entradas y Salidas Internacionales 2010. URL <http://anda.inec.gob.ec/anda/index.php/catalog/97> (Último acceso, Marzo 2012).
- INEC, 2011a. Índice de precios al consumo: canastas analíticas. URL http://www.inec.gob.ec/estadisticas/?option=com_content&view=article&id=135&Itemid=221 (Último acceso, Marzo 2012).
- INEC, 2011b. Censos Nacionales. Proyecciones de Población por Provincias, Cantones, Áreas y Grupos de Edad. Período 2001–2010. URL http://www.inec.gob.ec/inec/index.php?option=com_remository&func=fileinfo

- &id=940&Itemid=420&lang=es (Último acceso, Marzo 2012).
- Ishibuchi, H., Nozaki, K., Yamamoto, N., Tanaka, H., 1995. Selecting fuzzy if-then rules for classification problems using genetic algorithms. *IEEE Transactions on Fuzzy Systems* 3, 260–270.
- Izquierdo, M., Lacuesta, A., Vegas, R., 2009. Assimilation of immigrants in Spain: A longitudinal analysis. *Labour Economics* 16, 669–678.
- Jaakkola, T., 2000. Tutorial on variational approximation methods, in: Opper, M., Saad, D. (Eds.), *Advanced Mean Field Methods: Theory and Practice*, Neural Information Processing Series. The MIT Press, Cambridge, pp. 129–159.
- Jang, J.-S.R., 1993. ANFIS: Adaptive-network-based fuzzy inference system. *IEEE Transactions on Systems, Man and Cybernetics* 23, 665–685.
- Jennissen, R., 2003. Economic determinants of net international migration in Western Europe. *European Journal of Population* 19, 171–198.
- Jensen, F.V., Nielsen, T.D., 2007. *Bayesian networks and decision graphs*, 2nd ed, Information Science and Statistics. Springer New York, Nueva York.
- Jokisch, B., Pribilsky, J., 2002. The panic to leave: economic crisis and the “new emigration” from Ecuador. *International Migration* 40, 75–102.
- Kelton, W.D., 2009. Representing and generating uncertainty effectively, in: *Proceedings of the 2009 Winter Simulation Conference (WSC'09)*. IEEE, pp. 40–44.
- Khul, M.E., Lada, E.K., Steiger, N.M., Wagner, M.A., Wilson, J.R., 2008. Introduction to modeling and generating probabilistic input processes for simulation, in: *Proceedings of the 2008 Winter Simulation Conference (WSC'08)*. IEEE, pp. 48–61.
- Kiiveri, H., Speed, T.P., Carlin, J.B., 1984. Recursive causal models. *Journal of Australian Math Society* 36, 30–52.
- Kisioglu, P., Topcu, Y.I., 2011. Applying Bayesian Belief Network approach to customer churn analysis: A case study on the telecom industry of Turkey. *Expert Systems with Applications* 38, 7151–7157.
- Klir, G.J., 1991. Generalized information theory. *Fuzzy sets and systems* 40, 127–142.
- Klir, G.J., 2006. *Uncertainty and information foundations of generalized information theory*. Wiley-Interscience, Nueva Jersey.
- Kullback, S., Leibler, R.A., 1951. On information and sufficiency. *The Annals of Mathematical Statistics* 22, 79–86.
- Larrañaga, P., Karshenas, H., Bielza, C., Santana, R., 2013. A review on evolutionary algorithms in bayesian network learning and inference tasks. *Information Sciences* 233, 109–125.

- Larrea, C., 2004. Dolarización y desarrollo humano en Ecuador. *Iconos. Revista de Ciencias Sociales* 19, 43–53.
- Lauritzen, S., Spiegelhalter, D., 1988. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)* 50, 157–224.
- Lauritzen, S.L., 1996. *Graphical models*, Oxford Statistical Science. Clarendon Press, Oxford.
- Law, A.M., 2007. Statistical analysis of simulation output data: the practical state of the art, in: *Proceedings of the 2007 Winter Simulation Conference (WSC'07)*. IEEE, pp. 77–83.
- Lianos, T.P., 1997. Factors determining migrant remittances: the case of Greece. *International Migration Review* 31, 72–87.
- Liu, W.-Y., Yue, K., Su, J.-Y., Yao, Y., 2009. Probabilistic representation and approximate inference of type-2 fuzzy events in Bayesian networks with interval probability parameters. *Expert Systems with Applications* 36, 8076–8083.
- Lobo, F.G., Lima, C.F., Michalewicz, Z., 2007. *Parameter setting in evolutionary algorithms*, Studies in Computational Intelligence. Springer-Verlag Berlin Heidelberg, Berlin.
- López-Cevallos, D.F., Chi, C., 2012. Migration, remittances, and health care utilization in Ecuador. *Revista Panamericana de Salud Pública* 31, 9–16.
- Lucas, R.E., Stark, O., 1985. Motivations to remit: evidence from Botswana. *The Journal of Political Economy* 93, 901–918.
- Lueth, E., Ruiz-Arranz, M., 2007. Are workers' remittances a hedge against macroeconomic shocks? The case of Sri Lanka. *Asia-Pacific Development Journal* 14, 25–39.
- Makina, D., 2012. Migration and characteristics of remittance senders in South Africa. *International Migration* in press.
- Mamdani, E.H., Assilian, S., 1975. An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Man-Machine Studies* 7, 1–13.
- Massey, D.S., Arango, J., Hugo, G., Kouaouci, A., Pellegrino, A., Taylor, J.E., 2005. *Worlds in motion: understanding international migration at the end of the millennium*, International Studies in Demography. Clarendon Press, Oxford.
- McKenzie, D., 2005. Beyond remittances: the effects of migration on Mexican households, in: Schiff, M., Özden, Ç. (Eds.), *International Migration, Remittances and the Brain Drain*. Banco Mundial, Washington, pp. 123–147.
- Mckenzie, D., Rapoport, H., 2007. Network effects and the dynamics of migration and inequality: Theory and evidence from Mexico. *Journal of Development*

- Economics 84, 1–24.
- Merkle, L., Zimmermann, K.F., 1992. Savings, remittances, and return migration. *Economics Letters* 38, 77–81.
- Metropolis, N., Ulam, S., 1949. The Monte Carlo method. *Journal of the American Statistical Association* 44, 335–341.
- Mincer, J., 1958. Investment in human capital and personal income distribution. *Journal of Political Economy* 66, 281–302.
- Mohapatra, S., Ratha, D., Silwal, A., 2010. Outlook for remittance flows 2011-12: recovery after the crisis, but risks lie ahead (Migration and Development Brief No. 13). Banco Mundial, Washington.
- Moré, I., Echazarra, A., Halloufi, B., Petru, R., 2008. Cuantificación de las remesas enviadas por mujeres inmigrantes desde España (Documento de trabajo). Remesas.org, Madrid.
- Moreno, F.J., Bruquetas, M., 2011. Inmigración y estado de bienestar en España, Colección Estudios Sociales. Obra Social “La Caixa,” Barcelona.
- Nadkarni, S., Shenoy, P.P., 2004. A causal mapping approach to constructing bayesian networks. *Decision Support Systems* 38, 259–281.
- Nakayama, M.K., 2008. Statistical analysis of simulation output, in: *Proceedings of the 2008 Winter Simulation Conference (WSC’08)*. IEEE, pp. 62–72.
- Niimi, Y., Özden, Ç., Schiff, M., 2010. Remittances and the brain drain: skilled migrants do remit less. *Annales d’Economie et de Statistique* 123–141.
- OCDE, 2011. Social and welfare issues. Social Expenditure Database (SOCX). URL <http://www.oecd.org/social/soc/socialexpendituredatabasesocx.htm> (Último acceso, Marzo 2012).
- Osili, U.O., 2007. Remittances and savings from international migration: Theory and evidence using a matched sample. *Journal of Development Economics* 83, 446–465.
- Osyczka, A., 2002. Evolutionary algorithms for single and multicriteria optimization, *Studies in Fuzziness and Soft Computing*. Springer-Verlag New York, Nueva York.
- Park, H.-S., Cho, S.-B., 2012. Evolutionary attribute ordering in Bayesian networks for predicting the metabolic syndrome. *Expert Systems with Applications* 39, 4240–4249.
- Pearl, J., 1985. Bayesian networks: a model of self-activated memory for evidential reasoning, in: *Proceedings of the 7th Conference of the Cognitive Science Society*. Morgan Kaufmann Publishers, pp. 329–334.
- Pearl, J., 1988. Probabilistic reasoning in intelligent systems: networks of plausible

- inference, 2nd. ed.rev. ed. Morgan Kaufmann Publishers, San Francisco.
- Pearl, J., 2009. *Causality: models, reasoning and inference*, 2nd ed. Cambridge University Press, Nueva York.
- Pedersen, P.J., Pytlikova, M., Smith, N., 2008. Selection and network effects-migration flows into OECD countries 1990–2000. *European Economic Review* 52, 1160–1186.
- Pedrycz, W., 1994. Why triangular membership functions? *Fuzzy sets and Systems* 64, 21–30.
- Pedrycz, W., Reformat, M., 1997. Rule-based modeling of nonlinear relationships. *IEEE Transactions on Fuzzy Systems* 5, 256–269.
- Penz, C.A., Flesch, C.A., Nassar, S.M., Flesch, R.C.C., de Oliveira, M.A., 2012. Fuzzy-bayesian network for refrigeration compressor performance prediction and test time reduction. *Expert Systems with Applications* 39, 4268–4273.
- Piore, M., 1979. *Birds of passage: migrant labor and industrial societies*. Cambridge University Press, Cambridge.
- Raju, G.V.S., Zhou, J., Kisner, R.A., 1991. Hierarchical fuzzy control. *International Journal of Control* 54, 1201–1216.
- Rapoport, H., Docquier, F., 2006. The economics of migrants' remittances, in: Kolm, S.C., Ythier, J.M. (Eds.), *Handbook on the Economics of Giving, Reciprocity and Altruism*, Handbooks in Economics. North-Holland Publishing Co., pp. 1135–1198.
- Ratha, D., Ayana, G., Silwal, A., 2012. Remittances to developing countries will surpass \$400 billion in 2012 (Migration and Development Brief No. 19). Banco Mundial, Washington.
- Ratha, D., De Prabal, K., Mohapatra, S., 2011. Shadow sovereign ratings for unrated developing countries. *World Development* 39, 295–307.
- Ratha, D., Mohapatra, S., 2010. Forecasting migrant remittances during the global financial crisis. *Migration Letters* 7, 202–203.
- Ratha, D., Mohapatra, S., Silwal, A., 2010. *Migration and remittances*. Factbook 2011, 2nd ed. Banco Mundial, Washington.
- Ratha, D., Shaw, W., 2007. South-South migration and remittances (Documento de trabajo No. 102). Banco Mundial, Washington.
- Raychaudhuri, S., 2008. Introduction to Monte Carlo simulation, in: *Proceedings of the 2008 Winter Simulation Conference (WSC'08)*. IEEE, pp. 91–100.
- Reher Sullivan, D.S., Requena, M., Rosero-Bixby, L., 2009. Ecuatorianos en España, in: *Las Múltiples Caras de La Inmigración En España*. Alianza editorial, Madrid, pp. 117–152.

- Remesas.org, 2011. ¿Cuánto cuesta una remesa? URL <http://www.remesas.org/precioremesas.html> (Último acceso, Marzo 2012).
- Renooij, S., Witteman, C., 1999. Talking probabilities: communicating probabilistic information with words and numbers. *International Journal of Approximate Reasoning* 22, 169–194.
- Rizzi, L., Bazzana, F., Kasabov, N., Fedrizzi, M., Erzegovesi, L., 2003. Simulation of ECB decisions and forecast of short term Euro rate with an adaptive fuzzy expert system. *European Journal of Operational Research* 145, 363–381.
- Russell, S.S., 1986. Remittances from international migration: A review in perspective. *World Development* 14, 677–696.
- Schaffer, J.D., 1985. Multiple objective optimization with vector evaluated genetic algorithms, in: *Proceedings of the 1st International Conference on Genetic Algorithms*. L. Erlbaum Associates Inc., pp. 93–100.
- Shannon, C., 1948. A mathematical theory of communication. *Bell System Technical Journal* 27, 379–423.
- Shepherd, D., Shi, F.K.C., 2006. Fuzzy modelling and estimation of economic relationships. *Computational Statistics & Data Analysis* 51, 417–433.
- Simón, H., Sanromá, E., Ramos, R., 2008. Labour segregation and immigrant and native-born wage distributions in Spain: an analysis using matched employer–employee data. *Spanish Economic Review* 10, 135–168.
- Spiegelhalter, D.J., Dawid, A.P., Lauritzen, S.L., Cowell, R.G., 1993. Bayesian analysis in expert systems. *Statistical Science* 8, 219–247.
- Stark, O., Bloom, D., 1985. The new economics of labor migration. *American Economic Review* 75, 173–78.
- Straubhaar, T., 1986. The determinants of workers' remittances: the case of Turkey. *Weltwirtschaftliches Archiv* 122, 728–740.
- Takagi, T., Sugeno, M., 1985. Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man and Cybernetics SMC-15*, 116–132.
- Taylor, E.J., 1999. The new economics of labour migration and the role of remittances in the migration process. *International Migration* 37, 63–88.
- Todaro, M.P., 1969. A model for labor migration and urban unemployment in less developed countries. *American Economic Review* 59, 138–143.
- Ulyses Balderas, J., Nath, H.K., 2008. Inflation and relative price variability in Mexico: the role of remittances. *Applied Economics Letters* 15, 181–185.
- Uusitalo, L., 2007. Advantages and challenges of Bayesian networks in environmental modelling. *Ecological Modelling* 203, 312–318.

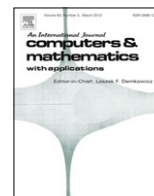
- Vasileva, K., 2012. Population and social conditions (Statistics in focus No. 31). Eurostat, Luxemburgo.
- Visscher, S., Lucas, P.J.F., Schurink, C.A.M., Bonten, M.J.M., 2009. Modelling treatment effects in a clinical Bayesian network using Boolean threshold functions. *Artificial Intelligence in Medicine* 46, 251–266.
- Walker, W.E., Harremoës, P., Rotmans, J., Van der Sluijs, J.P., Van Asselt, M.B.A., Janssen, P., Von Krauss, M.K., 2003. Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support. *Integrated Assessment* 4, 5–17.
- Weiss, Y., 2000. Correctness of local probability propagation in graphical models with loops. *Neural Comput.* 12, 1–41.
- Wermuth, N., Lauritzen, S.L., 1983. Graphical and recursive models for contingency tables. *Biometrika* 70, 537–552.
- White, K.P., Ingalls, R.G., 2009. Introduction to simulation, in: *Proceedings of the 2009 Winter Simulation Conference (WSC'09)*. IEEE, pp. 12–23.
- Wong, M.L., Guo, Y.Y., 2008. Learning bayesian networks from incomplete databases using a novel evolutionary algorithm. *Decision Support Systems* 45, 368–383.
- Yang, D., 2008. Coping with Disaster: The Impact of Hurricanes on International Financial Flows, 1970-2002. *The B.E. Journal of Economic Analysis & Policy* 8.
- Yang, D., Choi, H., 2007. Are Remittances Insurance? Evidence from Rainfall Shocks in the Philippines. *World Bank Economic Review* 21, 219–248.
- Zadeh, L.A., 1965. Fuzzy sets. *Information and Control* 8, 338–353.
- Zadeh, L.A., 1975. Fuzzy logic and approximate reasoning. *Synthese* 30, 407–428.
- Zhang, N.L., Poole, D., 1994. A simple approach to Bayesian network computations.
- Zhou, A., Qu, B.-Y., Li, H., Zhao, S.-Z., Suganthan, P.N., Zhang, Q., 2011. Multiobjective evolutionary algorithms: a survey of the state of the art. *Swarm and Evolutionary Computation* 1, 32–49.
- Zhu, Y., Liu, D., Jia, H., 2011. A new evolutionary computation based approach for learning bayesian network. *Procedia Engineering* 15, 4026–4030.
- Zimmermann, H.-J., 2000. An application-oriented view of modeling uncertainty. *European Journal of Operational Research* 122, 190–198.
- Zuo, Y., Kita, E., 2012. Stock price forecast using Bayesian network. *Expert Systems with Applications* 39, 6729–6737.

ANEXO I. PUBLICACIONES



Contents lists available at SciVerse ScienceDirect

Computers and Mathematics with Applications

journal homepage: www.elsevier.com/locate/camwa

A Multi-Objective Evolutionary Algorithm for enhancing Bayesian Networks hybrid-based modeling

Carlos R. García-Alonso^a, Pilar Campoy-Muñoz^{b,*}, Melania Salazar Ordoñez^b^a Universidad LOYOLA Andalucía, Departement of Quantitative Methods, Campus of Córdoba, C/ Escritor Castilla Aguayo 4, 14004, Córdoba, Spain^b Universidad LOYOLA Andalucía, Department of Economics, Campus of Córdoba, C/ Escritor Castilla Aguayo 4, 14004, Córdoba, Spain

ARTICLE INFO

Keywords:

Multi-Objective Evolutionary Algorithms
Fuzzy inference
Bayesian networks
Monte-Carlo simulation
Remittances

ABSTRACT

Bayesian Networks are increasingly being used to model complex socio-economic systems by expert knowledge elicitation even when data is scarce or does not exist. In this paper, a Multi-Objective Evolutionary Algorithm (MOEA) is presented for assessing the parameters (input relevance/weights) of fuzzy dependence relationships in a Bayesian Network (BN). The MOEA was designed to include a hybrid model that combines Monte-Carlo simulation and fuzzy inference. The MOEA-based prototype assesses the input weights of fuzzy dependence relationships by learning from available output data. In socio-economic systems, the determination of how a specific input variable affects the expected results can be critical and it is still one of the most important challenges in Bayesian modeling. The MOEA was checked by estimating the migrant stock as a relevant variable in a BN model for forecasting remittances. For a specific year, results showed similar input weights than those given by economists but it is very computationally demanding. The proposed hybrid-approach is an efficient procedure to estimate output values in BN.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

Remittances have been defined as “transfers made by migrants who are employed and have lived, at least one year, in other economies” [1]. In the case of developing countries, where the number of migrants by 2010 exceeds 171 million people (3% of the population), remittances have increased significantly over the past two decades, from US\$60,000 MM in 1990 to US\$325,000 MM in 2010 [2]. These financial flows amount to 2% of the GDP for developing countries and in some cases outnumber 25% of the GDP [3]. Remittance size, its counter-cyclical nature and their resilience have turned them into the second largest source of foreign currency for the developing world. Consequently, the sustainability debt of those countries is stronger, making their access to international capital markets also easier [3]. However, empirical studies have showed that remittances can also have negative effects on labor supply, inflation and real exchange rate, and mixed effects on the economic growth of the recipient economies [4].

In recipient economies, the estimation of the macroeconomic effects of remittances is critical. In this context, there have been some attempts to estimate remittance flows by using econometric techniques [5,6]. These studies highlighted that the lack of information and the poor quality of available data are the main obstacles for accurate forecasting. In addition, the economic literature about remittances reveals the complex nature of this phenomenon, in which a plethora of macroeconomic and microeconomic variables are involved [7]. On this basis, García-Alonso et al. [8] dealt with a remittance estimation problem by combining Bayesian Networks (BN), Monte Carlo simulation (MCS) and Fuzzy logic (FL).

* Corresponding author. Tel.: +34 957 222 100; fax: +34 957 222 181.
E-mail address: mpcampoy@uloyola.es (P. Campoy-Muñoz).

Expert-knowledge elicitation and learning from data can be used for stochastic assessment of BN. The former is a widely used alternative to develop BN when data is scarce or does not exist, but this process may be costly and time-consuming [9]. When data is available, several approaches can be used to develop BN, including structural equations [10] and learning algorithms [11]. Structural equations need prior expert knowledge to define the basic structure of the BN and then this method calculates the relevance of variables and their relationships according to a strong set of initial hypotheses [12]. Learning algorithms are devoted to discovering both the BN structure [13–15] and its parameters [16–18] from data. However, these algorithms can be very sensitive to the initial setting chosen by researchers and large amount of data are required [19]. In addition, given the complexity and stochastic nature of complex systems, algorithms have to operate in huge search spaces and they could be easily trapped into the numerous sub-optimal solutions [19]. To overcome some of these drawbacks, meta-heuristic approaches have been applied, such as evolutionary algorithms [20,21], evolutionary programming [22,23] and ant colony optimization [24,25]. Finally hybrid-approaches, combining expert knowledge and automated learning, have also being designed and developed to guide and speed up the process of learning efficient BN structures and parameters [26,27].

In this paper, the approach developed by [8] is enhanced by including a Multi-Objective Evolutionary Algorithm (MOEA). In [8], the relative importance (weight) of each input variable in its BN fuzzy dependence relationship (DR) is always defined by the experts as a random variable. The aim of this paper is to estimate these weights automatically when DR output values are known (secondary information). The determination of these weights can be seen as an Optimization Problem (OP), in which the combination of weights that optimizes a predefined fitness function could be found by using a bio-inspired searching procedure. Given an expert-based BN structure, including both probabilistic and fuzzy DR, the MOEA calculates input weights for the fuzzy ones knowing all or some output values. Thus, MOEA contributes to the evaluation of fuzzy DR and lets effective forecasting by learning the relative input relevance. Therefore, the hybrid-based approach offers a new strategy for stochastic evaluation of BN by combining the expert knowledge, Monte-Carlo simulation and Fuzzy inference with bio-inspired learning from data procedures.

This paper is organized as follows. Section 2 briefly describes the hybrid-based approach for stochastic assessment of BN, focusing on the role of input variable weights in fuzzy DR. Section 3 explains the OP regarding to input weight determination and offers a detailed description of the MOEA. In Section 4, the complete hybrid-based approach is used to estimate the migrant stock, as a relevant variable in the remittances model, and empirical results from a real case are shown. Finally, Section 5 concludes and discusses the future work.

2. The methodological framework

2.1. Bayesian Networks

Bayesian Networks (BN) have rapidly become a leading tool for modeling complex systems under uncertainty by combining graph theory and probability theory [10]. BNs are graphical representations of conditional dependence relationships among stochastic variables. In a BN each node is a stochastic variable and when a causal relationship between two nodes exists it is represented by an arrow. A BN can have a very complex structure including bidirectional arrows and cycles. In a directed acyclic graph G , that is a special type of BN, each node represents a random variable $X = [X_1, \dots, X_q]$ in D_X and arcs encode direct conditional dependence relationships between variables $X_{h_i} \rightarrow X_i$, where X_{h_i} is the parent of X_i and, in turn, X_i is the descendant of X_{h_i} . In G no cycles are permitted. Given the parents of the discrete variable X_i , denoted by pa_i , its conditional distribution is defined by $P(X_i = x_i) = p(x_i|pa_i)$. Thus, the joint probability distribution of G is:

$$p(x_1, x_1, \dots, x_n) = \prod_i p(x_i|pa_i). \quad (1)$$

Thus, a BN can fully describe qualitative and quantitative aspects of the real-world problem under study. Given the very frequent lack of data on socio-economic systems, BNs have to be developed mainly through expert knowledge. In these cases, experts and/or selected literature are consulted about the structure of the BN: nodes and arcs. According to [10], the structure of a graph G can be considered a BN when (i) it includes all the relevant variables and (ii) their relationships have a causal nature (Markovian compatibility). The stochastic evaluation of a BN can be intractable when dealing with complex systems because of the increasing number of variables and their relationships. However, given the BN structure, Monte-Carlo simulation and fuzzy logic can be considered an alternative to assess BN in complex socio-economic systems [8].

2.2. The dependence relationship structure

A BN can have two types of Dependence Relationships (DR): the probabilistic DR (denoted by f) and the fuzzy DR (denoted by u). Both DRs relate input variables (parental nodes) – always random variables- X_i with their corresponding output variables (descendant nodes) – consequently, random variables- Y describing the causal behavior of the system under study. A probabilistic DR $Y = f(X_i)$ usually has an algebraic structure that can be easily managed using, for example, Monte-Carlo simulation. On the other hand, a fuzzy DR $Y = u(X_i)$ is a set R of fuzzy “IF...THEN” standard rules ($d = 1, 2, \dots, c$):

$$R : \text{ALSO } \left[\text{IF AND } \left(x_i \in X_i \text{ isr } A_{id} \right) \text{ THEN } y_d \in Y_d \text{ isr } B_d \right] \quad (2)$$

i being the number of input variables ($i = 1, 2, \dots, n$), isr the semantic connector (equivalent to “is”), A_{id} the membership functions (type I fuzzy sets or semantic labels) for the inputs X_i in each rule d and, finally, B_d the membership functions for the output Y_d (in this case, each rule d has only one output y_d).

The total number of fuzzy rules r_{tot} in the domain under study is an exponential function of the number of membership functions e selected for the inputs and the number of fuzzy DR h [28]:

$$r_{tot} = he^n. \tag{3}$$

This number can be enormous and, due to that, the specific definition of each rule is completely impossible when experts are involved. The key question in resolving fuzzy DR [29] is to make both (i) the design of appropriate fuzzy rules and (ii) the output value y assessment completely automatic once a specific set of rules r in R ($r \in R$) are instantiated by input values x_i .

2.3. The simulation model

When inputs are considered as root variables (they have not parental) in the BN, Monte-Carlo Simulation (MCS) generates artificial data series according to their statistical distributions, always selected by experts and/or obtained from literature. Therefore, in both f and u DR input values x_i can come from MCS and/or from other DR (when they are descendants in other DR). When BN variables are continuous, the simulation model has a steady-state behavior being the error-based stopping criterion critical [30].

2.4. The fuzzy inference engine

A Fuzzy Inference Engine (FIE) was incorporated into the MCS engine to deal with the problem of evaluating fuzzy DR. FIE designs, instantiates and evaluates automatically only the rules needed, reducing the computer time. To do this, experts are asked about: (i) the basic structure of the DR, that is, their inputs and outputs, and (ii) some additional parameters [29]. These parameters include some guidelines about how the computer has to design and interpret the rules once they are instantiated by the input values. The instantiation process is completely automatic once the input values are known.

FIE calculates the membership function B_d for each output y_d by means of a weighted average (4). The w_{id} indicates the relevance of each input x_i with respect to the output y_d . These weights are determined at random by MCS according to predefined expert-based statistical distributions. Previously, the input values x_i have been fuzzified for the output y_d using type I fuzzy sets. The resulting membership functions for the input values A_{id} are used to calculate the final membership function for the corresponding output B_d .

$$B_d = \text{round} \left[\left(\frac{\sum_{i=1}^n w_{id} A_{id}}{\sum_{i=1}^n w_{id}} \right) + Q + D \right]. \tag{4}$$

Two parameters, D and Q , are also incorporated in (4) to include expert knowledge in obtaining B_d . Q is an integer that allows the computer to take into consideration the potential imprecise influence of inputs in each output defined by the experts. The parameter D , also an integer, allows the computer to interpret expert-based nuances about the fulfilment degree fd of the rules involved in each DR (fulfilment nuances). As a result the B_d will be biased (shift to the right/left) thus reflecting more accurately the behavior of the system according to the existent expert knowledge. The parameters Q and D are calculated based on two properties of the DR: its type and its intensity [29] described by experts and translated by system designers.

The structure of the FIE is summarized as follows:

- Step 1: Input values x_i , generated by MCS and/or other DR, are fuzzified. Their membership functions A_{id} are identified and their corresponding membership degrees $\mu_{A_{id}}(x_i)$ are calculated for the output y_d ($i = 1, 2, \dots, n$ and $d = 1, 2, \dots, c$).
- Step 2: Input weights w_{id} ($i = 1, 2, \dots, n$ and $d = 1, 2, \dots, c$) are determined by MCS.
- Step 3: For each A_{id} combination that instantiated a specific fuzzy rule, the FIE determines the fulfilment probability fd according to the DR intensity selected by the experts for the corresponding DR.
- Step 4: If fd is lower than a pseudo-random number generated by the computer, then the distortion D is randomly determined within an automatically generated range of integers $[-v, v]$ where v depends on fd value: the lower the fd , the wider the range.
- Step 5: For each A_{id} combination that instantiated a specific fuzzy rule, the FIE determines B_d as if the neutral DR type ($Q = 0$) had been selected.
- Step 6: According to the real DR type chosen by experts (it can be the neutral one, so Q can be equal to 0) and the B_d previously determined (step 5, neutral type option), Q is then calculated.
- Step 7: Taking into consideration D and Q values, the final membership function B_d for the output y_d is calculated for each A_{id} .
- Step 8: The resulting output value (crisp value) y_d is determined using the product-sum-gravity with superposition method [31].

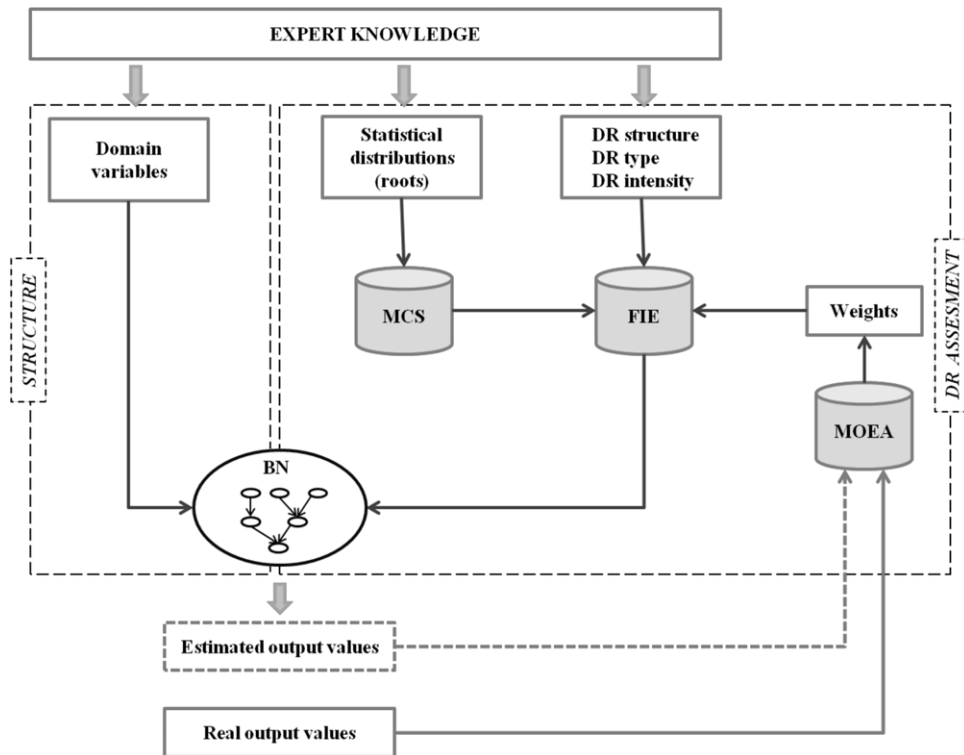


Fig. 1. Hybrid-based approach for modeling complex system under uncertainty.

3. Weight determination process

3.1. The multi-objective optimization problem

Under the foregoing approach, experts should decide at least three parameters or a group of parameters: the DR type (for calculating Q), the DR intensity (for calculating D) and the weights—relevance, step 2- w_{id} for each input x_i on the corresponding output y_d (for calculating B_d). While the two former ones depend directly on expert knowledge and are usually not controversial, the latter ones—weights and/or relevance—can be very arguable because they depend on the methodological approach and/or environmental conditions. From an experimental point of view, it could be interesting: (i) to determine which w_{id} combination or combinations could adjust the expected output value better and (ii) if this combination or set of combinations are stable in different geographic areas and/or throughout the time span. Then, the Pareto-based efficient solutions \mathbf{w}_e^* obtained ($e = 1, 2, \dots, z$ being z the number of total efficient solutions) could be compared to expert opinions and selected literature.

The \mathbf{W}^* , set of efficient solutions or weight combinations, can be obtained by the optimization of, at least, two objectives: (i) minimize the Mean Squared Error (MSE), between the estimated output \hat{y}_d and its real value y_d , and (ii) minimize the inner error (E) of the simulation process calculated according to Nakayama [30]. Once there is more than one non-linear objective, trade-offs between objectives are expected and, due to that, a set—not a unique optimum—of efficient solutions (feasible and non-dominated: the Pareto set) should be obtained. In our theoretical framework greater trade-offs are expected at the beginning of the evolutionary process because the natural evolution of \bar{y}_d .

Evolutionary algorithms (EA) are a suitable tool for optimizing non-linear functions and, therefore, to approximate the Pareto set [32]. The population-based strategy of EA and its simple selection strategy made them capable to approximate the whole Pareto set in a single run, instead of performing several separate runs as deterministic approaches do [33]. In addition, EA can deal with potential non-convex and discontinuous Pareto sets [33]. Therefore, an evolutionary algorithm has been designed and incorporated to include the previous methodological approach (Fig. 1) for solving the multi-objective optimization problem associated to the determination of input weights in fuzzy DR.

3.2. The Multi-Objective Evolutionary Algorithm

An EA is a meta-heuristic search algorithm based on biological principles of genetic and evolution. Since the prior work of Schaffer [34], there has been an increasing scientific interest in applying EAs to deal with multi-objective optimization problems [33]. Consequently, several approaches have been developed including the well-known plain aggregation

approach, the population-based non-Pareto approach, the Pareto-based approach and Niche induction approach [35], as well as other more recent developments described in detail in [33].

In our MOEA a solution j is a vector of input weights \mathbf{w}_j ($j = 1, 2, \dots, s$, being s the number of solutions in the population) which satisfies the following constraint:

$$\sum_{i=1}^n w_{ij} = 1 \quad (i = 1, 2, \dots, n, \text{ being } n \text{ the number of inputs in a specific fuzzy DR}). \tag{5}$$

In a specific generation g , the algorithm analyzed \mathbf{w}_{jg} , s being constant for all the generations. The \mathbf{w}_{jg} vector is the result of genetic operators on \mathbf{w}_{jg-1} till the stopping criterion is satisfied (the MOEA converges). A population \mathbf{W}_g is a set of s solutions for a specific generation g .

$$\mathbf{W}_g = [(w_{11}, w_{21}, \dots, w_{n1}), \dots, (w_{1s}, w_{2s}, \dots, w_{ns})] \quad (g = 1, 2, \dots, s). \tag{6}$$

Each solution in a generation g is analyzed by using two objectives: the Mean Square Error (MSE) and the inner error of the simulation process (E) calculated according to [30]. These objectives are calculated by a Monte-Carlo simulation engine that analyzes the corresponding statistical distributions selected for the inputs ($i = 1, \dots, n$). For each simulation k ($k = 1, \dots, m$) the value of the output y_{jgk} is calculated. Although the simulation process is steady-state, in order to simplify the validation process, the number of simulations m is considered constant for transforming it into a transient model [30]. According to this, the MSE, first objective, can be calculated:

$$f_{jg}^1 : \text{Min} \left[\frac{1}{m-1} \sum_{k=1}^m (y_{jgk} - y)^2 \right] \tag{7}$$

y being the output real value to be estimated. As stated above m is a constant for each solution j in each generation g . Therefore the previous can be simplified to:

$$f_{jg}^1 : \text{Min} \left[\sum_{k=1}^m (y_{jgk} - y)^2 \right]. \tag{8}$$

The second objective is the Nakayama error (E). Bootstrapping techniques are needed to calculate this value in order to avoid undesirable bias in variance determination. In a terminating simulation model, the number m of simulations is divided in a predefined number of batches b , usually: $10 \leq b \leq 30$ [30].

$$\bar{y}_{jgb} = \frac{1}{n_b} \sum_{k=n_b(b-1)+1}^{bn_b} y_{jgk} \tag{9}$$

for each batch b the mean \bar{y}_{jgb} of the obtained y_{jgk} values is calculated (n_b being the number of output values in each batch—usually identical). Then the grand mean \bar{y}_{jg} is calculated:

$$\bar{y}_{jg} = \frac{1}{b} \sum_{l=1}^b \bar{y}_{jgl}. \tag{10}$$

The standard deviation σ_{jg} can be easily calculated by:

$$\sigma_{jg} = \sqrt{\frac{1}{b-1} \sum_{l=1}^b (y_{jgl} - \bar{y}_{jgl})^2} \tag{11}$$

and, finally, the error E according to [30] is:

$$E_{jg} = \frac{t_{b-1, 1-\alpha/2} \sigma_{jg}}{\sqrt{b}} \tag{12}$$

$t_{b-1, 1-\alpha/2}$ being the $\alpha/2$ Student- t with $b - 1$ degrees of freedom. Therefore, the second objective function in our MOEA is:

$$f_{jg}^2 : \text{Min} (E_{jg}). \tag{13}$$

For each solution j in a generation g then we have objective values f_{jg}^1, f_{jg}^2 . For approximating the Pareto set, the fitness function proposed for the MOEA is based on the weighted-sum objective (WO) approach [36]:

$$\begin{aligned} \text{Min FF}_{jg} &= p_1 \tilde{f}_{jg}^1 + p_2 \tilde{f}_{jg}^2 \\ p_1 + p_2 &= 1 \\ p_1, p_2 &\geq 0 \end{aligned} \tag{14}$$

$\tilde{f}_{jg}^1, \tilde{f}_{jg}^2$ being the standardized $[0.1, 0.9]$ objective values and p_1 and p_2 their corresponding weights. These weights do not mean any kind of relevance at all. In order to avoid non-desired early convergence to a local optimum, p_1 and p_2 must be selected at random once the initial population of solutions has been designed. (14) is a constrain non-linear system with only one objective FF_{jg} that has to be optimized. The MOEA can easily incorporate other types of fitness functions like: SPEA2, ranking selection, fuzzy evaluation of weighted objectives.

When all the fitness functions (14) for each solution \mathbf{w}_{jg} are calculated, a stopping criterion is applied. This criterion evaluates the algorithm aptitude by the calculation of Mean Square Error of the fitness function $E_{FF_{jg}}$, which lets the algorithm identify the efficient solution set $\mathbf{W}_g^* = [\mathbf{w}_{1g}^*, \mathbf{w}_{2g}^*, \dots, \mathbf{w}_{zg}^*]$, $e = 1, 2, \dots, z$ being the number of efficient solutions in each generation g . When $E_{FF_{jg}}$ is repetitively lower than a predefined value e_{FF} or the maximum number of generations g_{\max} is reached, the algorithm stops and the solutions are saved in a solution pool SP .

If the stopping criterion is not fulfilled, genetics operators are systematically applied on \mathbf{W}_g for improving the fitness function of the next generation \mathbf{W}_{g+1} . A selection operator based on a standard binary tournament with replacement [32] is used, keeping the population size constant. Both weak and strict Pareto optimality criteria are employed during this process [32]. The dominated solutions \mathbf{w}_{jg} can be randomly replaced by new feasible solutions in order to increase the diversity in the search space [37]. An elitism operator is applied to a predefined number of the best solutions obtained. Standard crossover has no sense in this procedure because the genes in each solution are weights that follow the constraint (5), due to that, mutation probability has been increased. The mutation operator, which replaces one or more weights in a parent solution with other one/s, is performed at random. Both tournament and mutation operators are performed when the corresponding pseudo-random number is lower than the predefined tournament probability (p_T) or mutation probability (p_M). Unfeasible solutions, that is, those where the constraint (5) is not fulfilled, are detected and mutated by the MOEA till the feasibility is reached.

The Fig. 2 summarizes the MOEA procedure for determining the weights \mathbf{w}_j of fuzzy DR.

4. Experimental study

4.1. The remittances model and the migrant stock for a developed country

The remittances model proposed in [38] aims to estimate the remittance flows from a developed country to a developing one throughout the time span. This model is an expert-based BN (Fig. 3(a)), which includes relevant macroeconomic and microeconomic variables of these financial flows according to the New Economics of Labor Migration, Neoclassical Economic Theory and Wage Curve Approach. The model includes 20 random variables and 9 DRs, of which 3 are probabilistic DRs and 6 are fuzzy DRs. To solve it, 31 triangular membership functions were selected for each input/output variable in each fuzzy DR. Taking into account these values, the total number of fuzzy rules in the system under study was 1.4×10^{23} .

In this paper we focus our attention on a sub-network (BN_1), which describes the migrant stock in the developed country Nm_t (Fig. 3(b)) throughout the time span t . Two reasons have motivated the election of BN_1 for checking MOEA. First, Nm_t is a very relevant variable for estimating remittance flows as it represents the number of potential remitters [6]. Second, data about Nm_t is usually available (population census), and hence the results obtained \mathbf{w}_e^* can be compared both for several countries and throughout the time span.

From an economic point of view, the migrant stock Nm_t [39] can be calculated as the product of the population in the developing country PO_t and its propensity to migrate PM_t to the developed country. These variables define a probabilistic DR (DR_1):

$$Nm_t = f(PO_t, PM_t) = PO_t \cdot PM_t. \quad (15)$$

In turn, PM_t is defined by fuzzy DR (DR_2). Thus, PM_t is determined by (i) the social benefits BD_t [40] derived from public welfare policies in the developed country; (ii) the expected wage differential DW_t [41] between the countries involved; and finally, (iii) the migration cost I_t [42] from the developing to the developed country.

$$PM_t = u(BD_t, DW_t, I_t). \quad (16)$$

4.2. An illustrative example: the Ecuadorian migrant stock in Spain

In order to check the proposed MOEA, we have selected a real case study: the stock of Ecuadorian migrants in Spain during the year 2000, so $t = 1$. Table 1 shows the expert-based characteristics of BN_1 input variables: its statistical distribution (StD) and its influence on the corresponding output variable (orientation), which can be positive (+) – the greater the input value, the greater the output value, or negative (–) – the greater the input value, the lower the output value. Both DR type (Q) and DR intensity (D) are neutral, it means that there is no distortion in the inference process ($Q = 0$ and $D = 0$). Therefore, according to (3) the number of rules to be designed and evaluated is 29,971.

MOEA parameters were: (i) the population had 50 solutions with three weights according to (16), (ii) for each solution MCS ran 500 simulations in order to calculate the corresponding output value, (iii) for the fitness function FF_{jg} , the weights were $p_1 = p_2 = 0.5$, (iv) with respect to the stopping criterion, e_{FF} was fixed at 1% of the mean of $E_{FF_{jg}}$ for, at least,

- 1: **MOEA procedure**
- 2: Create the external solution file (*SP*)
- 3: **For** $seed=1$ to $seed_{max}$ **do** → being $seed_{max}$ a predefined number of different initial populations (seeds) \mathbf{W}_{seed} selected at random
- 4: Initial population design, \mathbf{W}_{seed} → Selected at random, each solution \mathbf{w}_j ($j=1, \dots, s$; s is the number of solutions in the population) in \mathbf{W}_{seed} satisfies $\sum_{j=1}^n w_{ij} = 1$
(n is the number of inputs in a specific fuzzy DR in a BN)
- 5: **for** $g=1$ to g_{max} **do** → being g_{max} a predefined number of generations
- 6: **OUTPUT/FITNESS CALCULATION procedure**
- 7: **For** $j=1$ to s **do**
- 8: **For** $k=1$ to m **do** → being m the maximum number of simulations (if the statistical error for y_{jgk} is repetitively lower than a predefined value the Monte-Carlo engine stops)
- 9: Calculate y_{jgk} → Monte-Carlo simulation and fuzzy inference engines according to the BN design
- 10: **end for**
- 11: Calculate $f_{jg}^1 : \frac{1}{m-1} \sum_{k=1}^m (y_{jgk} - y)^2$ and $\tilde{f}_{jg}^2 : E_{jg} = \frac{t_{b-1, 1-\alpha/2} \sigma_{jg}}{\sqrt{b}}$ → \mathbf{w}_j ($j=1, \dots, s$)
- 12: Calculate $FF_{jg} : p_1 \tilde{f}_{jg}^1 + p_2 \tilde{f}_{jg}^2$ → \mathbf{w}_j ($j=1, \dots, s$), p_1 and p_2 selected at random for each \mathbf{W}_{seed}
- 13: Calculate $E_{FF_{jg}}$ → $j=1, \dots, s$; Mean Square Error of the FF_{jg}
- 14: **end for**
- 15: **end procedure**
- 16: Calculate E_{FF_g} → Mean Square Error of the FF_{jg} ($j=1, \dots, s$)
- 17: **If** repetitively $E_{FF_g} \leq e_{FF}$ **go** to 23 → Stopping criterion
- 18: Selection of the best solutions w_j^* ($j=1, \dots, s$) → *Elitism*: $Min(p_1 \tilde{f}_{jg}^1 + p_2 \tilde{f}_{jg}^2)$
- 19: Preserve all non-dominated solutions in *SP*
- 20: **For** $j=1$ to s **do**
- 21: **If** $rand < p_T$ **do** binary tournament with replacement → p_T probability of being selected for a tournament
- 22: **end for**
- 23: **end for**
- 24: **end for**
- 25: **end procedure**

Fig. 2. MOEA pseudocode.

Table 1
Expert-based characteristics for input variables.

Variable ^a	Unit	DR	StD	Orientation	Q	D
PO _t	Inhabitants	DR ₁	U [7.25, 8.8]	N.A.	N.A.	N.A.
BD _t	Euros per inhabitant	DR ₂	U [3.70, 4.90]	+	Neutral	Neutral
DW _t	Times	DR ₂	U [1800, 3600]	+	Neutral	Neutral
I _t	Euros per migrant	DR ₂	U [1500, 5000]	-	Neutral	Neutral

^a "t" is the year analyzed.

10 consecutive generations, (v) elitism operator selected the two best solutions w_j^* in each generation g , (vi) the tournament probability was fixed at $p_T = 0.1$ and the mutation probability at $p_M = 0.1$ and, finally, (vii) no maximum number of generation g_{max} was established.

4.3. Results

Table 2 shows the basics statistics for the FF and each objective separately. The standardized values are low and then MOEA converges after $g = 44$ generations. Fig. 4 shows how the FF values are lower and closer to the mean of FF values as the number of generations increases. As results, the mean values for the input weights of the DR₂ are shown in Table 3. According to that, migration cost I_t from Ecuador to Spain is the most relevant variable to determine the propensity to migrate PM_t.

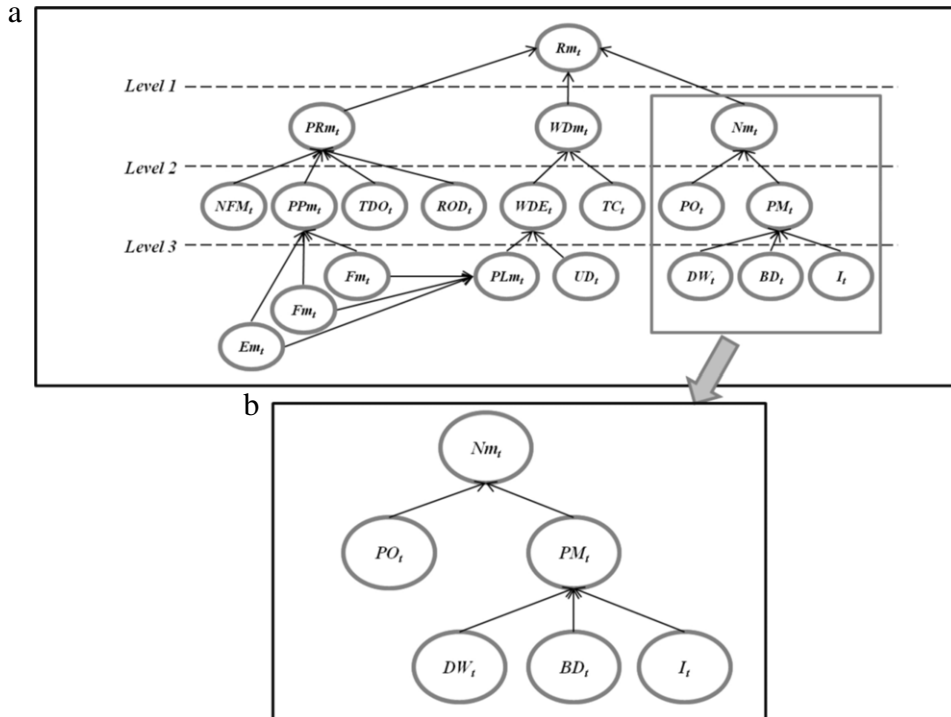


Fig. 3. (a) Remittances model and (b) Migrant stock model (BN₁).

Table 2

Basic statistics of standardized objectives [0.1, 0.9] in the external non-dominated solution file.

	Mean	SD ^a	Max	Min	#g
Min MSE	0.176	0.023	0.246	0.142	44
Min E	0.176	0.023	0.246	0.142	44
WO	0.201	0.021	0.248	0.166	44

^a Standard Deviation (SD).

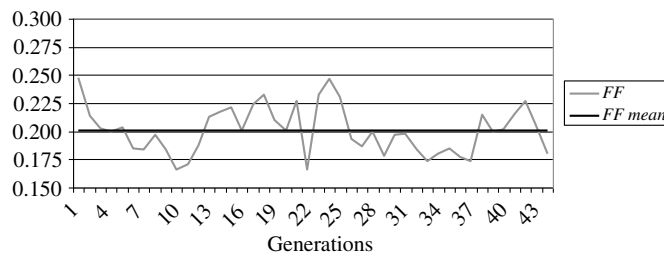


Fig. 4. Fitness function evolution.

The next most relevant variable is the wage differential DW_t between Spain and Ecuador. Finally, social benefits BD_t have the lowest influence on Ecuadorian propensity to migrate to Spain. These results are quite similar to those obtained in [38] in terms of order of importance of each input variable in the fuzzy DR, but the numerical values are different. It can be due to the fact that in [38] only one expert-based statistical distribution is selected for each weight throughout the time span and the MOEA has been carried out only for the first year of the period analyzed in [38].

Taking into account those weights, the mean estimated value for the stock of Ecuadorian migrants in Spain by 2000 was 122,903 (real value according to census was 124,887). Table 4 reflects that the estimated values were close to the real value by means of the MSE in percentage. Fig. 5 also displays how estimated values for Ecuadorian migrant stock are closer to its real value as the number of generations increases. On the other hand, the simulation model was very stable as the Nakayama error indicates [30].

Table 3
Basic statistics of each input weight in the external non-dominated solution file.

	Mean	SD ^a	Max ^b	Min ^b	#g
w_{BD_1}	0.235	0.034	0.336 (0.272)	0.182 (0.151)	44
w_{DW_1}	0.340	0.035	0.426 (0.257)	0.283 (0.205)	44
w_{I_1}	0.426	0.057	0.524 (0.240)	0.318 (0.233)	44

^a Standard Deviation (SD).
^b (SD) for the corresponding value.

Table 4
Basis statistics for output value Nm_t and the simulation model.

	Mean	SD ^a	%MSE ^b	%E ^c	#simulations
Nm_t	122 903.206	10 986.138	9.087	2.749	1.1×10^6

^a Standard Deviation (SD).
^b Mean Squared Error of output values, expressed as percentage of mean Nm_t .
^c Nakayama error as percentage of mean Nm_t .

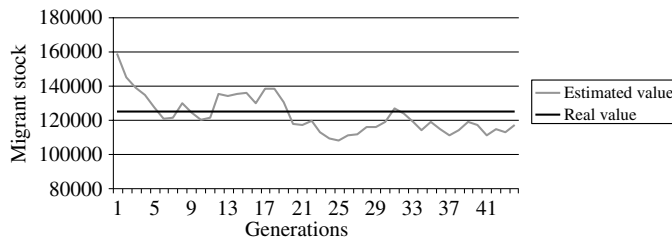


Fig. 5. Ecuadorian stock population in Spain during year 2000: Estimated values vs. Real values.

5. Conclusions

A hybrid-based approach is presented in this paper aiming to provide a new strategy for stochastic assessment of BN by combining expert knowledge and automated learning from available data, sometimes scarce. Given an expert-based structure of BN, the MCS was used to estimate the values of BN input variables throughout the time span. An inference engine was used to evaluate fuzzy DR and thus, expert knowledge is included into the simulation model making it possible to design all the rules needed and manage them in an automatic way. The MOEA aimed to reduce the number of parameters that expert must determine to design and evaluate these fuzzy rules. To do this, MOEA learns input weights (relevance) in the corresponding fuzzy DR from available output data.

The proposed approach is successfully applied to estimate the stock of Ecuadorian migrants in Spain by 2000, as a relevant variable in the BN remittances model that estimate those financial flows from Spain to Ecuador [38]. In addition, the input weights calculated by MOEA exhibit the same order of relevance that is previously given by experts. Since the selection of input weights is critical in order to draw the characteristics of the complex system on, MOEA introduces a higher precision into the output assessment procedure and hence it will allow the model to fit real situations better. In addition, it is worth emphasizing the fact that the proposed approach can be applied not only to estimate remittance flows but also to any complex system under uncertainty, in which expert will be able to define a dependence relationship structure among domain variables and some guidelines about their behavior.

Works are in progress to extend the MOEA procedure to calculate input weights throughout the time span. From a technical point of view, it can be easily managed by introducing new constraints into the fitness function. However, the challenging point is to determine how each input weight increases/decreases/keeps constant between periods throughout the time span. The input weights change over the time, but these changes are usually slight when socio-economic variables are involved, except in cases of revolutions, economic downturns, natural disasters, etc. Thus, the new constraints to be added and especially their corresponding parameters must be carefully chosen to capture the slight changes, but also the shocks on those variables.

Acknowledgements

This work was partially subsidized by the Spanish Inter-Ministerial Commission of Science and Technology under Project TIN2011-22794, the European Regional Development fund, and the “Junta de Andalucía”, Spain, under Project P08-TIC-374.

References

- [1] International Monetary Fund, Balance of Payments Statistics Yearbook 2010, International Monetary Fund, Washington, 2010.
- [2] World Bank, Migrations and Remittances Factbook 2011, World Bank, Washington, 2011.
- [3] D. Ratha, S. Mohapatra, A. Silwal, Migration and Development Brief 12, World Bank, Washington, 2010.
- [4] R. Chami, A. Barajas, T. Cosimano, C. Fullenkamp, M. Gapen, P. Montiel, Macroeconomic Consequences of Remittances, International Monetary Fund Occasional Paper No. 259, 2008.
- [5] S. Jimenez-Martín, N. Jorgensen, J.M. Labeaga, The Volume and Geography of Remittances from the EU, European Commission, Madrid, 2007.
- [6] S. Mohapatra, D. Ratha, Forecasting migrant remittances during the global financial crisis, *Migration Letters* 7 (2) (2010) 203–213.
- [7] J. Carling, The determinants of migrant remittances, *Oxford Review of Economic Policy* 24 (2008) 581–598.
- [8] C.R. García-Alonso, E. Arenas-Arroyo, G.M. Pérez-Alcalá, A macro-economic model to forecast remittances based on Monte-Carlo simulation and artificial intelligence, *Expert Systems with Applications* 39 (2012) 7929–7937.
- [9] J. Pitchforth, K. Mengersen, A proposed validation framework for expert elicited Bayesian Networks, *Expert Systems with Applications* 40 (2013) 162–167.
- [10] J. Pearl, *Causality: Models, Reasoning, and Inference*, second ed., Cambridge University Press, Cambridge, 2009.
- [11] R. Neapolitan, *Learning Bayesian Networks*, Prentice Hall, New Jersey, 2003.
- [12] A. Darwiche, *Modelling and Reasoning with Bayesian Network*, Cambridge University Press, Cambridge, 2009.
- [13] L.J. Yan, N. Cercone, Bayesian network modeling for evolutionary genetic structures, *Computers and Mathematics with Applications* 59 (2010) 2541–2551.
- [14] D. Heckerman, A tutorial on learning with Bayesian Networks, Tech. Rep., Microsoft Research, Redmond, Washington, 1995. URL: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.56.1431>.
- [15] J. Cheng, R. Greiner, J. Kelly, D. Bell, W. Liu, Learning Bayesian Networks from data: an information-theory based approach, *Artificial Intelligence* 137 (2002) 43–90.
- [16] S.T. Lauritzen, The EM algorithm for graphical association models with missing data, *Computational Statistics and Data Analysis* 19 (1995) 191–201.
- [17] M. Ramoni, P. Sebastiani, Parameter estimation in Bayesian Networks from incomplete databases, *Intelligent Data Analysis* 2 (1998) 139–160.
- [18] A. Feelders, L. van der Gaag, Learning Bayesian network parameters under order constraints, *International Journal of Approximate Reasoning* 42 (2006) 37–53.
- [19] S.H. Chen, C.A. Pollino, Good practice in Bayesian network modelling, *Environmental Modelling & Software* 37 (2012) 134–145.
- [20] D.Y. Liu, F. Wang, Y.N. Lu, W.X. Xue, S.X. Wang, Research on learning Bayesian network structure based on genetic algorithms, *Journal of Computer Research and Development* 38 (8) (2001) 916–922.
- [21] M.L. Wong, Y.Y. Guo, Learning Bayesian Networks from incomplete databases using a novel evolutionary algorithm, *Decision Support Systems* 45 (2008) 368–383.
- [22] M.M. Morales, R.G. Dominguez, N.C. Ramirez, A.G. Hernandez, J.L.J. Andrade, A method based on genetic algorithms and fuzzy logic to induce Bayesian Networks, in: *Proceedings of the Fifth Mexican International Conference in Computer Science*, 2004, pp. 176–180.
- [23] X. Li, S.M. Yuan, X. He, Learning Bayesian Networks structures based on extending evolutionary programming in: *Proceedings of 2004 International Conference on Machine Learning and Cybernetics*, 3, 2004, pp. 1594–1598.
- [24] L.M. Campos, J.M. Fernández-Luna, J.A. Gámez, J.M. Puerta, Ant colony optimization for learning Bayesian Networks, *International Journal of Approximate Reasoning* 31 (2002) 291–311.
- [25] J. Ji, R. Hu, H. Zhang, C. Liu, A hybrid method for learning Bayesian Networks based on ant colony optimization, *Applied Soft Computing* 11 (2011) 3373–3384.
- [26] W. Liao, Q. Ji, Learning Bayesian network parameters under incomplete data with domain knowledge, *Pattern Recognition* 42 (2009) 3046–3056.
- [27] L.M. Campos, J.G. Castellano, Bayesian network learning algorithms using structural restrictions, *International Journal of Approximate Reasoning* 45 (2) (2007) 233–254.
- [28] O. Cerdón, F. Herrera, I. Zwir, Fuzzy modeling by hierarchically built fuzzy rule bases, *International Journal of Approximate Reasoning* 27 (2003) 61–93.
- [29] C. García-Alonso, Dealing with complexity in large scale and structured fuzzy systems, in: *Proceedings of the Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, vol. 5, 2008, pp. 299–305.
- [30] M. Nakayama, Statistical analysis of simulation, in: *Proceedings of the 2008 Winter Simulation Conference*, 2008, pp. 62–72.
- [31] E. Cox, *Fuzzy Modeling and Genetic Algorithms for Data-Mining and Exploration*, Elsevier, San Francisco, 2005.
- [32] C.A. Coello-Coello, G.B. Lamont, D.A. van Veldhuizen, *Evolutionary Algorithms for Solving Multi-Objective Problems*, second ed., Springer, New York, 2007.
- [33] A. Zhou, B. Qu, H. Li, S. Zhao, P.N. Suganthan, Q. Zhang, Multiobjective evolutionary algorithms: a survey of the state of the art, *Swarm and Evolutionary Computation* 1 (2011) 32–49.
- [34] J.D. Schaffer, Multiple objective optimization with vector evaluated genetic algorithms, in: *Proceedings of the First International Conference on Genetic Algorithms*, 1985, pp. 93–100.
- [35] C.M. Fonseca, P.J. Fleming, An overview of evolutionary algorithms in multiobjective optimization, *Evolutionary Computation* 3 (1) (1995) 1–16.
- [36] I. Das, J. Dennis, A closer look at drawbacks of minimizing weighted sums of objectives for Pareto set generation in multicriteria optimization problems, *Structural Optimization* 14 (1) (1994) 63–69.
- [37] Z. Michalewicz, M. Schoenauer, Evolutionary algorithms for constrained parameter optimization problems, *Evolutionary Computation* 4 (1) (1996) 1–32.
- [38] P. Campoy-Muñoz, M. Salazar-Ordoñez, C.R. García-Alonso, Remittance flows estimation hybridizing a multilevel fuzzy system and a simulation model, in: *Proceedings of the Ninth International Conference on Fuzzy Systems and Knowledge Discovery*, 2012, pp. 138–142.
- [39] J. Hagen-Zanker, M. Siegel, The determinants of remittances: a review of the literature, *MGSoG Working Paper No. 003*, 2007.
- [40] G.J. Borjas, Immigration and Welfare Magnets, *Journal of Labor Economics* 17 (1999) 607–637.
- [41] M.P. Todaro, A model for labor migration and Urban unemployment in less developed countries, *American Economic Review* 59 (1) (1969) 138–148.
- [42] X. Clark, T.J. Hatton, J.G. Williamson, What explains emigration out of Latin America? *World Development* 32 (2004) 1871–1890.